Research Article

Evaluation of Speaker-Conditioned Target Speaker Extraction Algorithms for Hearing-Impaired Listeners

Trends in Hearing
Volume 29: 1–15

DOI: 10.1177/23312165251365802
journals.sagepub.com/home/tia
Sage

Ragini Sinha[1], Ann-Christin Scherer[1], Simon Doclo[1,2], Christian Rollwage[1] and Jan Rennies[1]

Abstract

Speaker-conditioned target speaker extraction algorithms aim at extracting the target speaker from a mixture of multiple speakers by using additional information about the target speaker. Previous studies have evaluated the performance of these algorithms using either instrumental measures or subjective assessments with normal-hearing listeners or with hearing-impaired listeners. Notably, a previous study employing a quasicausal algorithm reported significant intelligibility improvements for both normal-hearing and hearing-impaired listeners, while another study demonstrated that a fully causal algorithm could enhance speech intelligibility and reduce listening effort for normal-hearing listeners. Building on these findings, this study focuses on an in-depth subjective assessment of two fully causal deep neural network-based speaker-conditioned target speaker extraction algorithms with hearing-impaired listeners, both without hearing loss compensation (unaided) and with linear hearing loss compensation (aided). Three different subjective performance measurement methods were used to cover a broad range of listening conditions, namely paired comparison, speech recognition thresholds, and categorically scaled perceived listening effort. The subjective evaluation results with 15 hearing-impaired listeners showed that one algorithm significantly reduced listening effort and improved intelligibility compared to unprocessed stimuli and the other algorithm. The data also suggest that hearing-impaired listeners experience a greater benefit in terms of listening effort (for both male and female interfering speakers) and speech recognition thresholds, especially in the presence of female interfering speakers than normal-hearing listeners, and that hearing loss compensation (linear amplification) is not required to obtain an algorithm benefit.

Keywords

target speaker extraction, deep neural networks, hearing-impaired listeners, subjective evaluations, hearing aids

Received: December 29, 2024; revised: July 17, 2025; accepted: July 23, 2025

Introduction

The cocktail-party problem (Bronkhorst, 2015; Cherry, 1953) exemplifies a complex acoustic scenario in which individuals attempt to follow the conversation of a target speaker in the presence of multiple interfering speakers and background noise. Understanding the target speaker in such a multitalker scenario requires significantly more cognitive effort compared to a quiet setting. Even normal-hearing (NH) listeners often struggle to fully understand the target speaker under these conditions (Brungart et al., 2009; Kidd Jr et al., 2016). This becomes even more challenging for hearing-impaired (HI) listeners (Bacon et al., 1998; Reinten et al., 2021) due to peripheral hearing deficits that impair selective attention (Shinn-Cunningham & Best, 2008). Although the mechanisms underlying selective attention are not yet fully understood, one of the primary goals of speech processing research is to develop algorithms that can mimic these abilities and effectively extract the target speaker from a mixture. Such algorithms have significant potential for various real-world applications, including hearing aids and other assistive

[1]Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Germany
[2]Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

Corresponding Author:
Ragini Sinha, Fraunhofer-Institut für Digitale Medientechnologie IDMT, Marie-Curie-Straße 2, 26129, Oldenburg, Germany.
Email: ragini.sinha@idmt.fraunhofer.de

Data Availability Statement included at the end of the article

listening devices, such as smart earbuds and hearables that enhance conversational clarity in everyday environments, or remote microphones that transmit a target speaker in a classroom scenario.

With recent advancements in deep learning (Miikkulainen et al., 2024; Schmidhuber, 2015), this study focuses on deep neural network-based speaker extraction algorithms, commonly referred to as speaker-conditioned target speaker extraction (SC-TSE). In general, SC-TSE algorithms aim to directly extract the target speaker from the mixture utilizing auxiliary information about the target speaker (Žmolíková et al., 2023) (see detailed overview in the next section). Several SC-TSE algorithms have shown impressive performance when evaluated in terms of speech quality and intelligibility using commonly used objective measures, such as scale-invariant signal-to-distortion ratio (SI-SDR) (Le Roux et al., 2019), perceptual evaluation of speech quality (PESQ) (ITU-T, 2001), short-time objective intelligibility (STOI) (Taal et al., 2011), and word error rate (Wang & Chelba, 2003). However, despite their potential, there is a lack of studies investigating the benefits of the SC-TSE algorithms for HI listeners. This study aims to address this gap.

Recently, the performance of SC-TSE algorithms was subjectively evaluated by Sinha et al. (2023) and Thoidis and Goehring (2024). According to Thoidis and Goehring (2024), the evaluation was conducted using double-blind sentence recognition tests with both NH and HI listeners for a quasicausal SC-TSE algorithm for mixtures of one, two, and three speakers in the restaurant noise, comparing the performance against speech enhancement algorithms (without using auxiliary information about the target speaker). The results demonstrated that both NH and HI listeners benefited from the SC-TSE algorithm, with HI listeners experiencing a greater improvement compared to NH listeners. Despite these promising findings, the study's scope had notable limitations. Specifically, the language of the target speaker differed from that of the interfering speaker(s), and the signal-to-noise ratio (SNR) between target and interfering speaker(s) was fixed at 0 dB for all types of mixtures. When the target and interfering speakers speak different languages, extracting the target speaker from the mixture typically becomes easier for both human listeners (Freyman et al., 1999; Rhebergen et al., 2005) and speaker extraction algorithms (Wang et al., 2022). This is mainly due to the reduced informational masking and the presence of more distinct linguistic cues, compared to mixtures where both speakers use the same language. Besides, the SC-TSE algorithm used by Thoidis and Goehring (2024) was quasicausal, requiring ~ 5.75 ms of future information to operate effectively in real-time applications. According to Sinha et al. (2023), two different causal SC-TSE algorithms (referred to as Algo-1 and Algo-2), which do not require any future information, were evaluated with only NH listeners. The evaluation was conducted for mixtures of two and three speakers using three different behavioral evaluation methods, namely paired comparison, adaptive measurements of speech recognition thresholds (SRTs), and categorically scaled perceived listening effort. This evaluation covered a broad range of SNRs, with the target and interfering speakers using the same language. The findings revealed that Algo-2 provided significant benefits across all subjective measurement methods, while Algo-1 showed no improvement over the unprocessed mixture. Interestingly, both algorithms had shown considerable improvements in objective measures compared to the unprocessed mixture (Sinha et al., 2022; Wang et al., 2019). This discrepancy highlights that objective measures alone may not fully capture how humans perceive speech quality and intelligibility of the extracted target speaker. Such discrepancies are common, as objective measures are often based on simplified assumptions about distortions and artifacts, whereas SC-TSE algorithms can introduce complex, nonlinear artifacts that are perceptually noticeable to humans but not fully reflected in objective metrics. Therefore, subjective evaluations are essential to assess the potential and limitations of these algorithms before they can be effectively implemented in real-world applications like hearing aids.

In this study, we systematically investigate the benefits of SC-TSE algorithms for HI listeners using three different evaluation methods as used by Sinha et al. (2023) for the same language of target and interfering speakers, and for a broader range of SNRs, especially since different evaluation methods differ with respect to the applicable range of SNRs. Speech recognition tests for NH listeners are typically conducted at negative SNRs because ceiling performance is already reached below 0 dB SNR. However, in such challenging conditions, SC-TSE algorithms often struggle, as separating the target speaker from interfering speaker(s) becomes more difficult. At higher SNRs, where these algorithms tend to perform better, other measures like listening effort are better suited to evaluate speech perception. For HI listeners, speech-on-speech masking may result in SRTs of 0 dB or higher, potentially placing SC-TSE algorithms in a more favorable operational range. However, the impact of processing artifacts introduced by these algorithms on speech perception for HI listeners remains unknown. Additionally, it is currently unknown if and how a hearing loss compensation, as typically employed in hearing aids to restore audibility, interacts with SC-TSE algorithms. Addressing this gap is important for determining whether SC-TSE algorithms should be adapted differently for aided and unaided HI users, and whether algorithmic artifacts or distortions may be more noticeable at certain audibility levels. To address these gaps, this study focuses on the subjective performance evaluation of two SC-TSE algorithms with HI listeners, addressing the following research questions:

- Research Question 1: Do SC-TSE algorithms provide comparable or greater benefits for HI listeners compared to NH listeners, also at low SNRs (SNRs <0)

and high SNRs (SNRs >0), especially for the same language of the target and interfering speakers?
- Research Question 2: Does hearing loss compensation enhance the benefits of SC-TSE processing for HI listeners, or do listeners without hearing loss compensation experience similar benefits?

To answer these questions, we evaluated the potential of two SC-TSE algorithms to enhance the speech perception of the target speaker across a broad range of SNRs. The evaluations covered various acoustic conditions, including scenarios with one or two interfering speakers and with or without gender differences between the target speaker and the interfering speaker(s). Measurements were conducted both with (aided) and without (unaided) hearing loss compensation to assess the impact of compensation. Furthermore, we also compared the performance of the SC-TSE algorithms with HI listeners to the previous study performed with NH listeners (Sinha et al., 2023).

Target Speaker Extraction Algorithms

Target speaker extraction is closely related to both speech enhancement and blind source separation (BSS). While speech enhancement and target speaker extraction both aim to suppress undesired sources, BSS aims to estimate all individual sources from a mixture. The main distinction lies in their objectives: BSS separates all sources, whereas target speaker extraction isolates only the speaker of interest. Compared to speech enhancement, which typically addresses nonspeech interfering sources, target speaker extraction deals with interference from overlapping speech, making it especially relevant in multitalker scenarios.

Target speaker extraction aims at extracting the speaker of interest from a mixture of multiple speakers. One approach to achieve this is by first utilizing a BSS technique (Vincent et al., 2018) to separate all individual sources from the mixture and then select the target speaker utilizing a speaker selection module (Sinha et al., 2024). However, BSS techniques typically require the number of sources in the mixture to be known or estimated, which is not trivial in practice. Another approach involves using speech enhancement algorithms trained to enhance the dominant speaker in the mixture (Thoidis & Goehring, 2024). However, these algorithms often fail to generalize when the interfering speakers are at an equal or higher level than the target speaker, an issue frequently encountered in real-world scenarios.

An alternative approach is to use an SC-TSE algorithm (Žmolíková et al., 2023), which aims at directly extracting the target speaker from the mixture. SC-TSE algorithms typically require auxiliary information about the target speaker. The most commonly used types of auxiliary information include reference speech (Sinha et al., 2024; Wang et al., 2019; Xu et al., 2020; Žmolíková et al., 2019), visual cues (Ephrat et al., 2018), speech activity (Delcroix et al., 2021),

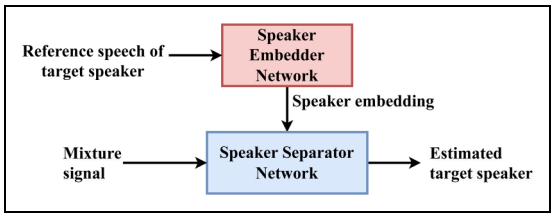

Figure 1. Block Diagram of Speaker-Conditioned Target Speaker Extraction (SC-TSE) Algorithm.

or directional information (Brendel et al., 2020; Gu et al., 2019) about the target speaker.

In this study, we consider two different SC-TSE algorithms that use the reference speech of the target speaker as auxiliary information (Delcroix et al., 2020; Ge et al., 2020; Sinha et al., 2024, 2022; Wang et al., 2019; Xu et al., 2020). The reference speech is a prerecorded utterance of the target speaker which is different from the utterance of the target speaker used in the mixture. Figure 1 depicts the block diagram of a typical SC-TSE algorithm, consisting of two networks: a speaker embedder network and a speaker separator network. The goal of the speaker embedder network is to generate a speaker embedding from the reference speech of the target speaker. The target speaker embedding represents condensed speech features of the target speaker which guides the separator network toward estimating the target speaker from the mixture.

For this study, we used the same algorithms (Algo-1 and Algo-2) previously evaluated by Sinha et al. (2023) with NH listeners. It should be noted that the primary aim of this study is not to compare the algorithms, but to investigate whether the performance trends observed in objective evaluation metrics align with the subjective outcomes in HI listeners, as was similarly investigated for NH listeners (Sinha et al., 2023). Inspired by Wang et al. (2019), Algo-1 employs separate training of the embedder and separator networks and estimates a real-valued mask in the short-time Fourier transform (STFT) domain to perform the speaker extraction. It should be noted that due to the use of a real-valued mask, Algo-1 does not estimate the phase of the target speaker in the STFT domain but uses the phase of the mixture. Algo-1 uses low complexity ResNet-gated recurrent units (ResNet-GRU) in the separator network and a pretrained long short-term memory network in the embedder network. For Algo-1, the total computational complexity in terms of number of multiplications and additions (MACs) is 4.1 G. Algo-2 (Sinha et al., 2022) employs joint training of the embedder and separator networks to perform the speaker extraction in the time domain, hence implicitly estimating the phase of the target speaker. Algo-2 uses a combination of a temporal convolutional network and a convolution-augmented transformer (conformer) in the speaker separator network and a ResNet in the embedder network. Due to the multihead self-attention mechanism in the conformer, the total computational

complexity required for Algo-2 is 16.9 G, which is approximately 4.12 times larger than Algo-1. Both Algo-1 and Algo-2 are causal, with an algorithmic latency of 32 ms for Algo-1, and 2.5 ms for Algo-2, for details see (Sinha et al., 2023).

Both algorithms were trained on the same dataset for mixtures of two speakers, mixtures of three speakers, and noisy mixtures of two speakers at a sampling rate of 16 kHz. For creating the training and validation sets of the two-speaker and three-speaker mixtures, we used the *si_tr_s* subset of the WSJ0 corpus (Hershey et al., 2016). We followed the same WSJ0-2mix dataset creation script, used in several baseline algorithms (Deng et al., 2021; Ge et al., 2020; Liu et al., 2023; Xu et al., 2020; Zhang et al., 2023). To create the mixtures of two speakers, two different speakers from the WSJ0 corpus (Hershey et al., 2016) were randomly chosen and mixed at an SNR between 0 and 5 dB, where the first speaker was considered as the target speaker and the second speaker as the interfering speaker. The SNR was randomly sampled between 0 and 5 dB rather than fixed at a specific value. A different utterance of the target speaker was randomly chosen as the reference speech of the target speaker to generate the speaker embedding. Similarly, the mixtures of three speakers were created where both interfering speakers had the same power, and the mixture with the target speaker was simulated at an SNR between 0 and 5 dB. The noisy mixtures of two speakers were created using the official WHAM corpus simulation scripts (Wichern et al., 2019), where the target and interfering speakers were selected from the WSJ0 corpus, and the noise was selected from the WHAM corpus. In total, we generated 47,926 utterances for training and 12,792 utterances for validation, using 101 speakers for training and 18 speakers for validation across all three mixture types. The training, validation, and evaluation sets were fully disjoint, with no overlap in speakers across the sets.

It should be noted that both algorithms were trained and validated using mixtures composed of English speech, while the subjective evaluations were conducted using German speech materials. Algo-1 was trained using the scale-invariant signal-to-distortion ratio (SI-SDR) loss (Luo & Mesgarani, 2019), while Algo-2 used a weighted combination of multi-scale SI-SDR loss for speaker separator and cross-entropy loss for speaker embedder network, as by Ge et al. (2020). Both algorithms were optimized with the ADAM optimizer for up to 150 epochs with early stopping.

Participants and Stimuli

Participants

Fifteen native German-speaking HI listeners (7 males, 8 females), aged 54–65 years, participated in the listening test experiments. Figure 2 shows the group mean and individual audiograms for the right and left ears. All participants underwent laboratory-based audiometric testing to confirm mild to moderate sensorineural hearing loss (Bisgaard et al., 2010) based on their pure-tone thresholds. The hearing loss was relatively symmetric (the difference in pure-tone average between the left and right ears was <10 dB for all participants, except one, who showed a difference of 20 dB). All participants received hourly compensation and gave informed consent for their participation in the experiments. The methods were approved by the ethics committee of the University of Oldenburg (protocol Drs.EK/2019/073-02).

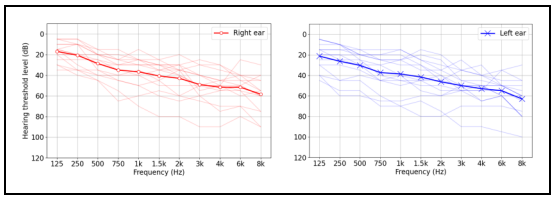

Figure 2. Individual and Group Mean Hearing Thresholds (in dB) for the Right and Left Ears for the Participants.

Stimuli and Equipment

The same stimuli as by Sinha et al. (2023) were used in this study. The target speaker stimuli consisted of German matrix sentences uttered by a fixed male speaker from the Oldenburg sentence test (OLSA) (Wagener et al., 1999). The reference speech of the target speaker was chosen from the German Göttingen sentence test (GÖSA) (Kollmeier & Wesselkamp, 1997) which consists of everyday sentences uttered by the same male speaker. Each OLSA sentence followed a fixed syntactical structure containing five words in the following order: name, verb, numeral, adjective, and object. For each word, 10 alternatives were available, which were randomly combined to generate syntactically correct but semantically unpredictable sentences. The interfering speech also consisted of matrix sentences uttered by one or two different speakers (either male or female), chosen from Hochmuth et al. (2018). Interfering speaker signals were generated by concatenating several sentences starting at a random position for each presentation. The relative level of the target speaker and the interfering speaker(s) was varied to produce different SNRs (see below). As the reference speech of the target speaker, several utterances of the target speaker from the GÖSA sentence test were concatenated to make a 10s-long utterance. This utterance was used to obtain a 256-dimensional target speaker embedding vector, which was consistently used across both algorithms during the evaluation. Although both SC-TSE algorithms were also trained with noisy mixtures of two speakers, no such stimuli were included in the experiments.

To familiarize the participants with the voice of the target speaker, each participant listened to an example of about 60*s* consisting of concatenated sentences uttered by the target speaker. These sentences were mixed with interfering speakers as in the experiments, but at a high SNR (between +5 and +10 dB) to ensure that the target speaker was much

louder than the interfering speakers. During the experiments, stimuli were presented diotically via Sennheiser HD650 headphones in sound-attenuated booths.

Subjective Evaluation Methods and Procedures

To assess the performance of both considered SC-TSE algorithms, paired comparisons (Parizet, 2002), speech intelligibility measurements (Wagener et al., 1999), and perceived listening effort scaling (Rennies et al., 2014) were utilized. These methods vary in terms of the outcome measure and the SNR range to which they are applicable. Paired comparisons were used to determine the preferences of participants between different versions of the same stimulus. An SNR = 0 dB was used because this test scenario is typically considered in instrumental evaluations of SC-TSE algorithms. Speech intelligibility was measured in terms of SRTs, that is, SNRs corresponding to 50% speech intelligibility. SNRs are typically negative at such low performance levels, at least for NH listeners (as observed in, e.g., Kidd Jr et al., 2016; Rennies et al., 2019; and Sinha et al., 2023). As observed by Sinha et al. (2023), the mean SRTs were −5.0 dB for the male interfering speakers and −13.0 dB for the female interfering speakers. Categorical listening effort scaling was used to assess the perceived effort that a participant needed to understand the target speaker. This method is typically measured over a broad range of SNRs.

In this study, we used either one or two interfering speakers in the unprocessed stimuli, depending on the specific evaluation method applied. Paired comparisons and perceived listening effort were measured for stimuli in which the target speaker was masked by either one or two interfering speakers, while SRTs were measured only for stimuli having two interfering speakers. We excluded SRT measurements with only one interfering speaker, as SRTs for such conditions are known to be extremely low (Rhebergen et al., 2005), that is, falling into SNR regions where algorithms are not expected to work well, nor where typical listening conditions would occur (Smeds et al., 2015). For all methods the signals were initially scaled such that the target speaker had the same level (65 dB SPL before hearing loss compensation) as the single or the combined interfering speaker(s), and then the target speaker was adapted to generate stimuli at different SNRs. In the processed conditions, these mixtures were processed by either Algo-1 or Algo-2, typically reducing the level of the interfering speaker(s) energy. In the aided conditions, hearing loss compensation was applied afterwards.

Each evaluation was conducted for three processing conditions (unprocessed, Algo-1, and Algo-2), two genders of interfering speakers (male and female), and different number of interfering speakers (depending on the evaluation method). In the paired comparison, all combinations of three pairwise processing conditions, one and two interfering speakers, and two genders were considered, resulting in 12 unique conditions. Each condition was repeated three times, leading to 36 trials per participant for both unaided and aided. For the SRT measurements, only two interfering speakers were considered, combined with three processing conditions and two genders, resulting in six unique conditions and approximately 120 trials per participant for both unaided and aided. The listening effort measurement included all combinations of three processing conditions, one and two interfering speakers, two genders, and five SNRs, resulting in 60 unique conditions. Each condition was repeated three times, leading to 180 trials per participant for both unaided and aided.

Paired Comparisons: In each trial, two versions of the same stimulus were presented to the participants, visually labeled as intervals A and B. The participants could toggle between these two versions as many times as they liked by clicking on the intervals. The stimuli were played in a loop, allowing participants to decide in which interval the target speaker was more intelligible. Participants were asked to rate on a six-point scale if one interval was "much easier" (German: "viel einfacher") (A+++/B+++), "clearly easier" ("deutlich einfacher") (A++/B++), or "easier" ("einfacher") (A+/B+) intelligible than the other. The middle category ($A = B$) on the rating scale was deliberately omitted in this study, requiring participants to select one interval in each trial. All three versions (unprocessed, stimuli processed by Algo-1, and stimuli processed by Algo-2) were compared with each other, while in each trial the assignment to intervals A and B was randomized. For each comparison, three repetitions were performed using different targets and interfering sentences. The outcomes of each comparison were analyzed in terms of the percentage of wins.

Speech Recognition Thresholds: SRTs were measured adaptively. On each trial, a mixture of the target speaker and two interfering speakers (processed or unprocessed) was presented once to the participants. The participants were asked to mark the recognized words on a word matrix shown on the screen before proceeding to the next trial. The level of the combined interfering speakers was fixed at 65 dB SPL (before hearing loss compensation), while the level of the target speaker was adjusted based on the participants responses in the preceding trial. If participants correctly identified three or more words out of five, the SNR was decreased, otherwise, the SNR was increased. The initial SNR was 5 dB and the step size was varied according to the procedure proposed by Brand and Hohmann (2002) to converge to the SRT. To avoid clipping and excessively loud stimuli, the maximum SNR was set to 20 dB. Each version of the stimuli (unprocessed, processed by Algo-1, and processed by Algo-2) was measured with a different list of 20 distinct sentences, presented in a random order. To reduce training effects, two training lists containing 20 target speaker sentences mixed with stationary noise were measured (Wagener et al., 1999) before conducting the actual SRT measurements.

Perceived Listening Effort: In each trial, a mixture of the target speaker and interfering speaker(s) (processed or unprocessed) was presented to the participants. The participants were asked to rate the perceived effort required to understand the target speaker on a 13-point scale ranging from "no effort" (German: "mühelos") corresponding to 1 effort scaling categorical unit (ESCU, see Figure 1 by Krueger et al., 2017) to "extreme effort" ("extrem anstrengend") (13 ESCU). An additional 14th category "only interfering speakers" ("nur Störsprecher") was included for trials in which participants could only hear the interfering speaker(s) (when an assessment of the effort related to listening to the target speaker could not be reasonably made). These assessments employed stimuli with predetermined SNRs as by Rennies et al. (2014). During each trial, the stimulus was played continuously in a loop until participants provided their rating, after which the next trial started. All SNRs and processing conditions were presented in a random order. The target and interfering speaker(s) were mixed at SNRs ranging from −10 to 15 dB, with a step size of 5 dB. The overall stimulus level was kept fixed at 65 dB SPL (before hearing loss compensation). For each combination of SNR, processing condition, and interfering speaker condition (one or two), the measurement was repeated three times using distinct sentences. The median value obtained from these repetitions was utilized as the assessment of an individual's perceived listening effort for that specific combination.

All measurements were performed for both unaided and aided conditions. For the unaided condition, no hearing loss compensation was provided, while for the aided condition, individualized amplification was provided to the participants. The stimuli were amplified by applying a linear gain according to the National Laboratories Revised Profound (NAL-RP) prescription (Dillon, 2012). For each participant, the amplification applied to the left and right ears was identical and calculated based on the average hearing threshold across both ears.

Results

Paired Comparisons

Figure 3 shows the percentage of wins from the paired comparison tests for both unaided (left column) and aided (right column) conditions. The top and middle panels compare unprocessed stimuli with stimuli processed by Algo-1 and Algo-2, while the bottom panels compare Algo-1 directly with Algo-2. Different hatches/colors represent different masking conditions (M/F: one male/female interfering speaker and MM/FF: two male/female interfering speakers). For both unaided and aided conditions, the data reveal a relatively similar pattern of ratings, where a clear preference for Algo-2 was observed compared to unprocessed stimuli and Algo-1. Stimuli processed by Algo-2 were favored in comparison to unprocessed stimuli in 100% of all comparisons, and in 98% for two male (MM) interfering speakers. The category "much easier" (+++) was given most often for one male/female and two female interfering speaker(s), while "clearly easier" (++) was given most often for two male interfering speakers. Similarly, in the direct comparison between the algorithms, participants frequently rated Algo-2 as "much easier" to understand than Algo-1. In neither unaided nor aided conditions, participants found any benefit of Algo-1 compared to unprocessed stimuli as most ratings were given to the middle categories of the rating scale, indicating that participants were uncertain about making a decision. These observations were supported by statistical analyses: ordinal values 0 to 5 were assigned to the six response categories, and one-sample Wilcoxon signed-rank tests were conducted for each combination of interferer type, comparison pair, and hearing loss compensation (24 tests in total) to test if the median values of the response distributions were significantly different from 2.5, that is, the hypothetical center of the scale. The significance level was Bonferroni-corrected for multiple comparisons. The tests indicated statistically significant differences for all comparisons including Algo-2, while none of the comparisons between unprocessed stimuli and stimuli processed by Algo-1 showed response distributions with median values that were significantly off-center.

Speech Recognition Thresholds

Figure 4 shows the measured averaged SRTs (top panels) and the corresponding improvements achieved by each algorithm compared to the unprocessed stimuli (bottom panels). For both unaided (left column) and aided (right column) conditions, mean SRTs were considerably lower for female interfering speakers than for male interfering speakers. Mean SRTs obtained for unprocessed stimuli were −5.1 dB (female) and 0.0 dB (male) for the unaided condition, and −5.9 dB (female) and −1.3 dB (male) for the aided condition. For both unaided and aided conditions, Algo-1 showed no benefits in mean SRTs (even an increase) compared to unprocessed stimuli, whereas Algo-2 showed considerable benefits for both female and male interfering speakers. For the unaided condition, Algo-2 achieved 1.2 dB (female) and 3.2 dB (male) lower SRTs. For the aided condition, the benefit was 1.6 dB (female) and 2.9 dB (male). For the aided condition, mean SRTs were negative for all three processing conditions (unprocessed, Algo-1, and Algo-2), indicating that participants were able to understand 50% of the target speaker even when the energy of the interfering speakers exceeded that of the target speaker.

It should be noted that, for the unaided condition, the SRT measurement of three participants were invalid for Algo-1 with male interfering speakers. This occurred because participants mistakenly followed one of the interfering speakers instead of the target speaker. As a result, the adaptive procedure kept increasing the SNR until the predefined maximum of 20 dB was reached. At three ceiling hits, the trial was aborted automatically.

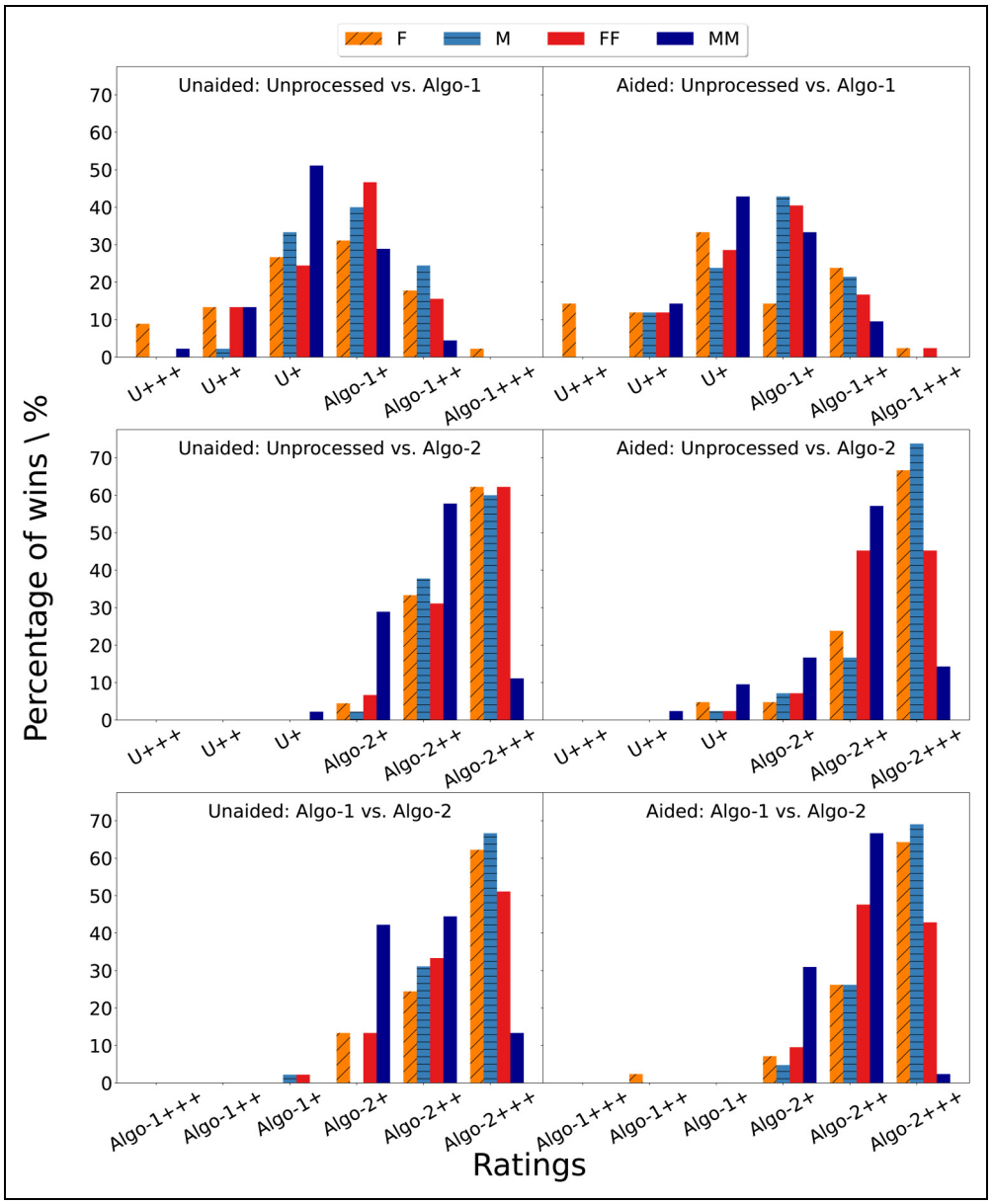

Figure 3. Percentage of Wins From the Paired Comparison Tests Obtained for Each Pair of the Three Processing Conditions (Unprocessed, Algo-1, and Algo-2) for Stimuli Having One (*F*/*M*) or Two (*FF*/*MM*) Interfering Speakers. The Left Column Represents the Ratings for the Unaided Conditions, and the Right Column the Ratings for the Aided Conditions. The Utilized SNR was 0 dB. The Reported Percentage of Wins Reflects the Proportion of Responses, Aggregated Across All Participants and All Three Repetitions. Abbreviations: M/F = One Male/Female Interfering Speaker; MM/FF = Two Male/Female Interfering Speakers; SNR = Signal-to-Noise Ratio.

Statistical analyses were performed using a linear mixed-effects model with the lme4 package in R software (Bates et al., 2015), which is well-suited for handling missing data (invalid data from the three participants were considered as missing data). Participants were treated as a random factor. We conducted a comprehensive diagnostic evaluation, including visual inspection of posterior predictions, linearity, homogeneity of variance, collinearity, influential observations, normality of residuals, and normality of random effects using the performance package in R (Lüdecke et al., 2021). Furthermore, we performed contrast analysis with Holm corrections using the model-based package (Makowski et al., 2020), with an alpha level of .05 for all tests. Visual inspections of the residuals of the linear mixed-effects model predicting the outcome of variable SRTs revealed a normal distribution.

A linear mixed-effects model was fitted to the measured SRTs to estimate the fixed effects of processing (unprocessed, Algo-1, and Algo-2), hearing loss compensation (unaided and aided), and the gender of interfering speakers (female and

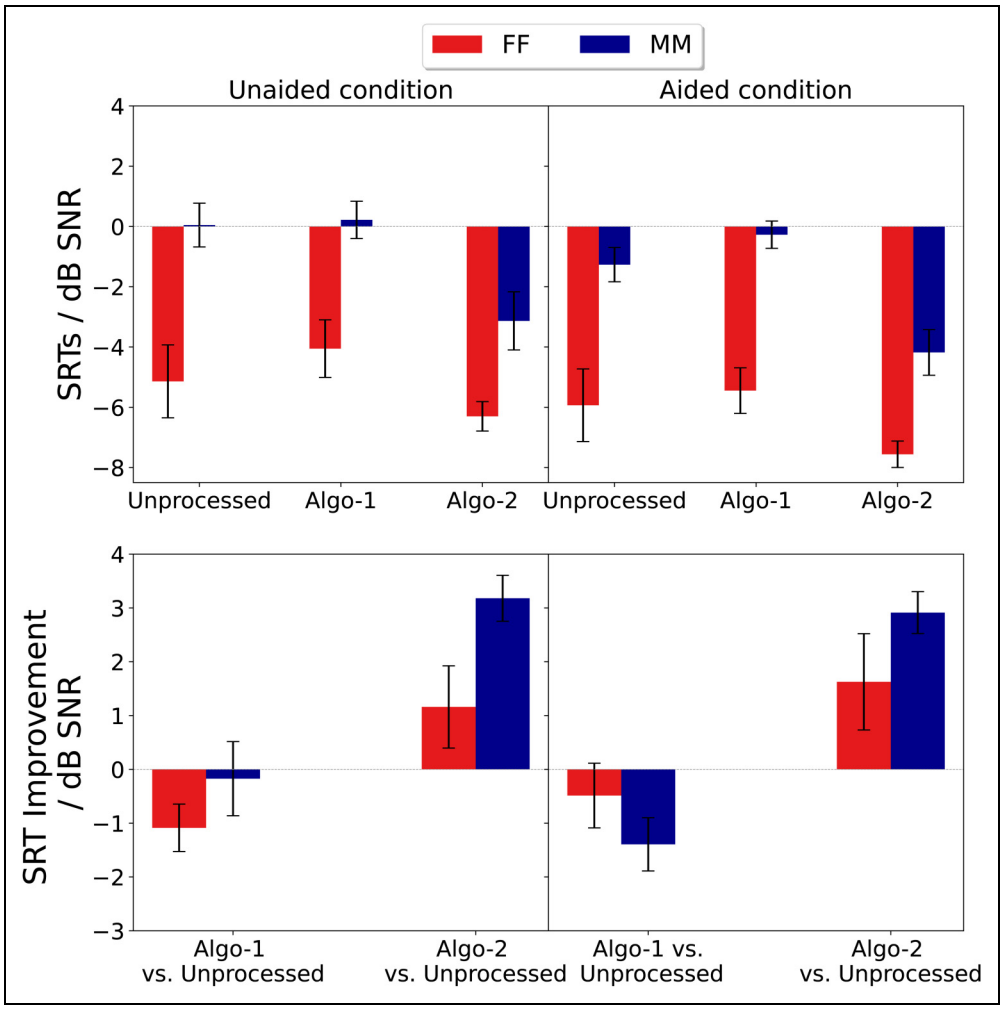

Figure 4. Speech Recognition Thresholds (SRTs) Averaged Across All Participants (Top) and Corresponding SRT Improvements (Bottom) Obtained for Stimuli Having Two Female (*FF*) or Male (*MM*) Interfering Speakers. The Left Column Represents SRTs and Corresponding Improvements for the Unaided Conditions, and the Right Column the Aided Conditions. Error Bars Represent the Standard Errors.

male), along with their two- and three-way interactions. An analysis of variance revealed significant main effects of processing, $F(2, 154) = 44.6, p < .001$, hearing loss compensation, $F(1, 154) = 16.2, p < .001$, and the gender of interfering speakers, $F(1, 154) = 267.7, p < .001$. Additionally, a significant two-way interaction was found between processing and the gender of interfering speakers, $F(2, 154) = 3.9, p = .021$. However, neither a statistically significant three-way interaction nor an interaction between hearing loss compensation and processing were found.[1]

Significant effects were further analyzed using contrast analysis to compare the three factors (see Table 1). Table 1 is divided into two parts. The first part presents the main effects of processing, hearing loss compensation, and the gender of interfering speakers. The second part presents the pairwise differences between all levels of processing for each gender of interfering speakers and the pairwise differences between all levels of gender for each processing. Only statistically significant differences are reported. The results revealed significant differences between unprocessed and Algo-2, as well as between Algo-1 and Algo-2, for both male and female interfering speakers. However, no significant difference was found between unprocessed and Algo-1. Additionally, for each processing condition (unprocessed, Algo-1, and Algo-2), a significant difference was observed between male and female interfering speakers.

Perceived Listening Effort

Figure 5 shows the median listening effort ratings across participants along with the corresponding benefits for one and two interfering speakers for both unaided and aided conditions as a function of SNR. The first three rows represent the listening effort ratings for unprocessed stimuli, Algo-1, and Algo-2, while the last row represents the listening effort benefits of both algorithms compared to unprocessed stimuli.

Table 1. Results of Contrast Analysis for Predicting the Differences in SRTs. Only Significant Differences Are Reported.

Part 1

Effect	Difference (dB SNR)	p Value
Male–female	4.3	<.001
Unaided–aided	1.1	<.001
Algo-1–Algo-2	2.9	<.001
Unprocessed–Algo-1	−0.7	.030
Unprocessed–Algo-2	2.2	<.001

Part 2

Effect	Interfering speakers	Difference (dB SNR)	p Value
Algo-1–Algo-2	Female	2.2	<.001
Algo-1–Algo-2	Male	3.7	<.001
Unprocessed–Algo-2	Female	1.4	.005
Unprocessed–Algo-2	Male	3.1	<.001
Effect	**Processing**	**Difference (dB SNR)**	**p Value**
Male–female	Algo-1	4.8	<.001
Male–female	Algo-2	3.3	<.001
Male–female	Unprocessed	4.9	<.001

In general, listening effort ratings systematically decreased with increasing SNR for both one and two interfering speakers (except for Algo-1 at high SNRs), and followed a similar pattern for unaided and aided condition. For unprocessed stimuli and low SNRs, the perceived effort was higher with two interfering speakers compared to one interfering speaker. For two interfering speakers, participants also rated the 14-th category "only interfering speakers" for male (unaided) and for female (aided) at the lowest SNR (−15 dB). Algo-1 showed a minimal reduction in listening effort at 5 dB SNR (one interfering speaker), but no reduction at other SNRs. At higher SNRs, it even increased the listening effort compared to unprocessed stimuli, likely due to artifacts such as residuals of interfering speakers that affect the overall quality and intelligibility of the processed signal. It can be observed that for both NH listeners (Sinha et al., 2023) and HI listeners, interfering speakers were most perceptible at the highest SNR for Algo-1. This may be due to residual interference and processing artifacts becoming more noticeable at higher SNRs, likely reflecting the limited ability of Algo-1 to generalize to SNRs outside its training range, leading to suboptimal suppression and more prominent artifacts. In contrast, Algo-2 reduced perceived listening effort at all considered SNRs for both one and two interfering speakers. Participants gave a median rating of "no effort" for one interfering speaker (male and female) at every SNR except at −10 dB, and for two female interfering speakers except at −10 and −5 dB. Overall, Algo-2 showed significant benefits at all considered SNRs compared to unprocessed stimuli for both one and two interfering speakers.

To investigate the effect of SNR and hearing loss compensation on listening effort ratings, we conducted a statistical analysis of the listening effort benefit provided by Algo-1 and Algo-2 compared to the unprocessed stimuli. A linear mixed-effects model was fitted to the listening effort benefit to estimate the fixed effects of processing benefits (Algo-1 vs. unprocessed, and Algo-2 vs. unprocessed), hearing loss compensation (unaided and aided), and SNRs, along with their two-way interactions. An analysis of variance revealed significant main effects for processing benefits, $F(1, 1412) = 789.8$, $p < .001$, and SNRs, $F(5, 1412) = 118.2$, $p < .001$, as well as a significant two-way interaction between processing benefits and SNRs, $F(5, 1412) = 23.1$, $p < .001$. However, no statistically significant main effect or two-way interactions including hearing loss compensation were observed. The significant effects and interactions were further analyzed using contrast analysis (see Table 2).

Table 2 is divided into two parts. Part 1 presents the main effect of processing benefit, while Part 2 presents the pairwise comparison of processing benefit for each SNR. Only statistically significant differences are reported. The main effects of SNR and the pairwise comparisons between all SNRs for each processing benefit type were also analyzed. The results revealed that, for each SNR, there was a significant difference in the benefits provided by Algo-1 and Algo-2. Additionally, significant differences were found between most SNR pairs, except for the following cases: (10 dB vs. −10 dB, 15 dB) for Algo-2; (−5 dB vs. −10 dB, 5 dB) for Algo-1; (−5 dB vs. 0 dB) for both Algo-1 and Algo-2; (−10 dB vs. 10 dB) for Algo-2; and (−10 dB vs. −5 dB) for Algo-1.

Discussion

Differences for Unprocessed Stimuli Between NH and HI

The present data revealed considerable differences in terms of SRTs and listening effort ratings between NH (see Figures 2

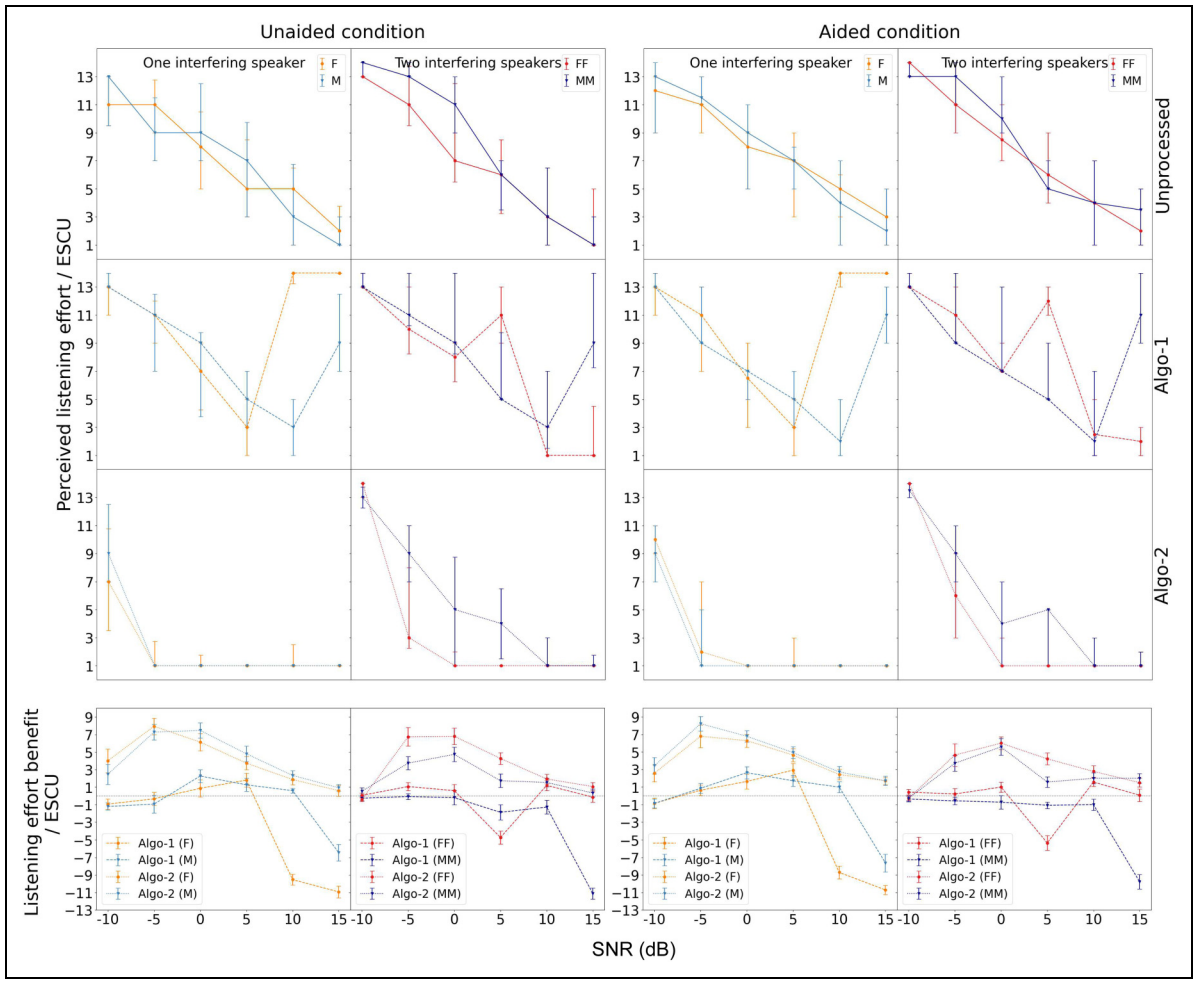

Figure 5. Median Perceived Listening Effort Ratings and Benefit Relative to Unprocessed Stimuli as a Function of SNR for Stimuli Having One (*F*/*M*) or Two (*FF*/*MM*) Interfering Speaker(s). The First Three Rows Represent the Listening Effort Ratings for Unprocessed Stimuli, Algo-1, and Algo-2, while the Last Row Represents the Listening Effort Benefit of Algo-1 and Algo-2 Compared to Unprocessed Stimuli. The Left Columns Represent the Unaided Condition, and the Right Columns the Aided Condition. Error Bars Represent the Interquartile Difference for the Perceived Listening Effort Ratings, and Standard Errors for the Listening Effort Benefit.

and 3 by Sinha et al., 2023) and HI listeners (see Figures 4 and 5) for the unprocessed stimuli.[2] SRTs of HI listeners were considerably higher than SRTs of NH listeners with a difference of 5 dB for male interfering speakers and 8 dB for female interfering speakers. Similar findings for SRT measurements with HI listeners were also reported by Kidd et al. (2019) in measurements employing similar matrix-type sentence material. The perceived listening effort ratings further indicated that HI listeners had to put more effort to understand the target speaker compared to NH listeners for both one and two interfering speakers, especially at lower SNRs. Moreover, unlike NH listeners (Sinha et al., 2023), HI listeners did not experience any reduction in effort at the lowest SNR (−10 dB) when the interfering speakers were of the opposite gender to the target speaker. Altogether, in line with previous studies (Hygge et al., 1992; Kidd et al., 2019; Shinn-Cunningham & Best, 2008), this study confirms that HI listeners struggle more in complex multitalker scenarios compared to NH listeners.

Algorithm Benefits for NH Versus HI

Similarly to NH listeners (Sinha et al., 2023), Algo-1 did not provide any improvement over unprocessed stimuli for HI listeners in terms of preference, speech intelligibility, or listening effort, while Algo-2 consistently demonstrated improvements over unprocessed stimuli for all considered evaluation methods. Notably, Algo-2 provided greater improvements for HI listeners compared to NH listeners for all evaluation methods. For NH listeners, Algo-2 improved mean SRTs (∼ 3 dB) only for male interfering speakers compared to the unprocessed stimuli, while for HI listeners, Algo-2 showed improvements for both male (∼ 3 dB) and female (∼ 1 dB) interfering speakers. This could be because of

Table 2. Results of Contrast Analysis for Predicting the Differences in Listening Effort Benefits. Only Significant Differences Are Reported.

Part 1			
Effect		Benefit Difference	*p* Value
Algo-1–Algo-2		−5.0	<.001
Part 2			
Effect	SNR (dB)	Benefit Difference	*p* Value
Algo-1–Algo-2	−10	−2.0	<.001
Algo-1–Algo-2	−5	−6.0	<.001
Algo-1–Algo-2	0	−5.2	<.001
Algo-1–Algo-2	5	−4.4	<.001
Algo-1–Algo-2	10	−4.2	<.001
Algo-1–Algo-2	15	−8.3	<.001

the much higher SRTs for unprocessed stimuli with female interfering speakers for HI listeners (−5 to −6 dB) compared to NH listeners (−13 dB). When comparing benefits between NH and HI listeners, Algo-2 showed a similar benefit for both NH and HI listeners for male interfering speaker with no statistically significant difference according to pairwise contrast analysis using a linear mixed-effects model. In contrast, for the female interfering speakers, significant differences to NH listeners of 4.7 dB (aided) and 4.2 dB (unaided) were observed (aided: 95% CI [2.75, 6.56], $p < .001$; unaided: 95% CI [2.28, 6.09], $p < .001$).

Algo-2 also showed a reduction in listening effort over a broad range of SNRs for both NH listeners (Sinha et al., 2023) and HI listeners for both one and two interfering speakers. However, the benefits were more pronounced for HI listeners with reductions of 7-8 ESCU compared to 4-5 ESCU for NH listeners at medium SNRs. Even at the lowest SNR (−10 dB) for one interfering speaker, the benefit for HI listeners (4-5 ESCU) was greater than for NH listeners (1-2 ESCU). Pairwise contrast analysis using a linear mixed-effects model between NH and HI listeners for the listening effort benefits also showed that HI listeners in both unaided and aided conditions showed significantly larger benefits compared to NH listeners at each SNR (all $p < .001$).

The paired comparison results showed a similar trend for both NH listeners (Sinha et al., 2023) and HI listeners. However, HI listeners displayed a stronger preference for Algo-2, with the majority of their ratings falling into the "much easier" category of the rating scale. Across all evaluation methods, Algo-2 provided greater benefits for HI listeners compared to NH listeners. This overall result aligns with findings from Thoidis and Goehring (2024), who observed similar outcomes in a double-blind sentence recognition test with a different SC-TSE algorithm.

We also performed Pearson correlation analyses to assess whether hearing loss severity, as measured by PTA4 (averaged across both ears), is related to the SRTs and listening effort benefits of Algo-2, analyzed separately for male and female interfering speakers. No significant correlations were observed between hearing loss severity and algorithm benefit across any condition, unaided or aided, with one or two interfering speakers, or at any SNR.

Impact of Algorithmic Artifacts

In general, SC-TSE algorithms may introduce artifacts in the processed signals, mainly depending on the SNR of the mixture, the number of interfering speakers and the used speaker separator network. Typical artifacts include distortions of the extracted target speaker and residual interference from the interfering speaker(s). As already mentioned, Algo-1 performs target speaker extraction in the STFT domain using a real-valued mask (hence using the mixture phase) with a relatively simple network architecture, whereas Algo-2 performs target speaker extraction in the time domain with a more complex network architecture. Several studies (Ge et al., 2020; Pandey & Wang, 2018; Xu et al., 2019) have shown that real-valued spectral masking-based approaches typically introduce more artifacts compared to time-domain approaches.

A significant impact of artifacts introduced by Algo-1 was also observed in HI listeners. Algo-1 not only failed to reduce listening effort compared to unprocessed stimuli but, in some cases even increased the required effort. Specifically, the listening effort was higher at 10 dB (female) and 15 dB (male) for one interfering speaker, and at 5 dB (female) and 15 dB (male) for two interfering speakers. These findings are inconsistent with objective assessments using PESQ (ITU-T, 2001) and STOI (Taal et al., 2011), where Algo-1 showed considerable improvement in target speaker quality and intelligibility (Wang et al., 2019). This discrepancy arises because Algo-1 introduces artifacts that significantly impair speech perception for human listeners, but which may not be fully reflected by objective evaluation metrics. Artifacts, such as residuals of the interfering speakers, may hinder the listener's ability to selectively attend to the target speaker during subjective evaluations, which is typically not captured by objective metrics. The artifacts introduced by Algo-1 had a more pronounced effect on HI listeners (without hearing loss compensation) than on NH listeners, as observed during SRT measurements. For male interfering speakers in the unaided condition, some HI listeners started to follow one of the interfering speakers instead of the target speaker, leading to invalid measurement data. For HI listeners, it is particularly challenging to focus on the target speaker when the mixture includes interfering speakers of the same gender. Algorithm artifacts, such as residuals of the interfering speaker(s) in the processed signal, exacerbate this difficulty. Without hearing loss compensation, HI listeners are more likely to miss crucial features of the target speaker's voice that help distinguish it from interfering speakers. The fact that no invalid adaptive tracks occurred when hearing loss compensation was applied may indicate that (partially) restoring audibility can help to extract information from stimuli containing processing artifacts.

Effects of Hearing Loss Compensation

Apart from the appearance of such invalid tracks during the SRT measurements, this study found no significant differences between measurements without (unaided) and with (aided) hearing loss compensation. Results from all three evaluation methods showed very similar patterns for both unaided and aided conditions. Even though SRTs for the aided condition (unprocessed, Algo-1, and Algo-2) were slightly lower than for the unaided condition, the difference was small (1.1 dB on average), and no significant interactions between hearing loss compensation and other factors were found. The statistical analysis for listening effort benefits also confirmed this. Similar observations were reported by Ohlenforst et al. (2017) for hearing loss compensation. Although linear amplification can make sounds audible for HI listeners in noisy environments, it does not significantly improve the ability to understand speech in multitalker scenarios. This study confirms that linear amplification alone is not sufficient to improve target speech perception (see results for unprocessed stimuli for all evaluation methods). Therefore, to solve the cocktail party problem, target speaker extraction algorithms such as the ones investigated in this study are required, as they can provide substantial benefits with or without hearing loss compensation. In other words, hearing loss compensation may not be needed for these speaker extraction algorithms to be beneficial, especially when the participant has mild to moderate hearing loss. We found the algorithmic benefit to be independent of amplification, which applies only to the group tested in this study and may differ for individuals with more severe hearing loss. One possibly interesting implication of this observation is that speaker extraction algorithms could be a valuable future asset of hearables that target a broader user range than people with known hearing loss. This target group could benefit from assistive listening comprising speaker extraction even when increased hearing thresholds play a minor role in their daily life or when they are not aware of audibility impairments. We also observed that speech audibility was not a critical factor in this study, as no substantial differences were found between the aided and unaided conditions.

One limitation of this study is that all participants had a relatively symmetric hearing loss; it is possible that the role of hearing loss compensation could also be different for the participants with asymmetric hearing loss, where both ears should be amplified differently. Additionally, all participants had mild to moderate hearing loss. Therefore, the effects of hearing loss compensation observed here may not generalize to individuals with more severe hearing loss, who may respond differently to SC-TSE algorithms. Future work will include a broader range of hearing loss severities to better understand how hearing loss compensation influences algorithmic performance across diverse listener profiles.

Limitations of Algorithms and Future Directions

The results of this study suggest that HI listeners could benefit significantly from SC-TSE algorithms if these are implemented in practical applications such as hearing aids. However, for an algorithm to be suitable for hearing aids, it needs to meet specific requirements. The algorithm needs to be capable of real-time processing, that is, have low computational complexity and an algorithmic latency of 10 ms or less (Agnew & Thornton, 2000; Bramsløw, 2010). Additionally, the algorithm needs to meet the hardware constraints of the device, balancing performance and power consumption. In this study, both evaluated SC-TSE algorithms are causal and in principle capable of real-time processing. Algo-2 demonstrated substantial benefits across all evaluation methods and meets the latency requirement (2.5 ms). However, it falls short in other critical aspects, such as memory size and computational complexity. Currently, the required total number of MACs for Algo-2 is 16.9 G. In future work, we will focus on making Algo-2 more suitable for hearable applications by reducing the required overall memory size and computational complexity, for example, by implementing a more efficient multi-head self-attention mechanism in the conformer. Additionally, we further plan to investigate the specific factors contributing to the performance differences between Algo-1 and Algo-2, along with a comprehensive study comparing the performance of both algorithms with their nonspeaker-conditioned versions. Furthermore, this study focuses solely on the impact of one and two interfering speakers in the mixture signal. The effects of other factors, such as background noise and reverberation, remain unexplored for both NH listeners and HI listeners, which needs further investigation in future work.

Conclusions

The following main conclusions can be drawn from this study:

- Similar to findings with NH listeners, Algo-2 demonstrated considerable benefits for all considered evaluation methods with HI listeners, while Algo-1 did not show any benefit compared to unprocessed stimuli. HI listeners experienced a greater reduction in listening effort at lower SNRs and a greater improvement in SRTs compared to NH listeners, particularly in the presence of female interfering speakers, while both groups exhibited similar SRT gains when the interfering speakers were male. This suggests that SC-TSE algorithms can be effective in enhancing target speech perception and reducing the perceived listening effort for HI listeners, potentially providing even greater benefits than for NH listeners.
- Artifacts introduced by SC-TSE algorithms, especially Algo-1, were observed to impact HI listeners, similarly as for NH listeners.

- Although susceptibility to processing artifacts may be increased in unaided conditions for some listeners, none of the evaluation methods showed a significant impact of hearing loss compensation. Results for both unaided and aided conditions were relatively similar for all considered evaluation methods, indicating that the benefit of SC-TSE algorithms does not depend on hearing loss compensation using linear amplifications.

Acknowledgments

The authors thank Jonathan Albert Göäwein for his valuable suggestions and insightful discussions regarding the statistical analysis of data.

ORCID iDs

Ragini Sinha https://orcid.org/0000-0001-7456-0121
Ann-Christin Scherer https://orcid.org/0009-0007-3538-447X
Simon Doclo https://orcid.org/0000-0002-3392-2381
Jan Rennies https://orcid.org/0000-0002-0291-7723

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The Oldenburg Branch for Hearing, Speech and Audio Technology HSA is funded in the program ≫ *Vorab* ≪ by the Lower Saxony Ministry of Science and Culture (MWK) and the Volkswagen Foundation for its further development. This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project ID 352015383—SFB 1330 A1 and B2 and Project ID 390895286—EXC 2177/1.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest for the research, authorship, and/or publication of this article.

Data Availability Statement

All data supporting this study are not publicly available due to privacy and ethical reasons. However, they can be made available from the corresponding author upon a reasonable request. The evaluation stimuli were used as the same stimuli used in Sinha et al. (2023).

Notes

1. We also explored using mean, median, and multiple data imputation approaches rather than treating the data as missing in the linear mixed-effects models to assess whether different handling of missing data would affect the analysis. The results remained consistent.
2. An interactive graphic to compare the results of each measurement method between NH and HI listeners can be found at: https://raginisinha.github.io/examples_german.github.io/.

References

Agnew, J., & Thornton, J. M. (2000). Just noticeable and objectionable group delays in digital hearing aids. *Journal of the American Academy of Audiology*, *11*(06), 330–336. https://doi.org/10.1055/s-0042-1748062

Bacon, S. P., Opie, J. M., & Montoya, D. Y. (1998). The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds. *Journal of Speech, Language, and Hearing Research*, *41*(3), 549–563. https://doi.org/10.1044/jslhr.4103.549

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(i01), 1–48. https://doi.org/10.18637/jss.v067.i01

Bisgaard, N., Vlaming, M. S., & Dahlquist, M. (2010). Standard audiograms for the IEC 60118-15 measurement procedure. *Trends in Amplification*, *14*(2), 113–120. https://doi.org/10.1177/1084713810379609

Bramsløw, L. (2010). Preferred signal path delay and high-pass cutoff in open fittings. *International Journal of Audiology*, *49*(9), 634–644. https://doi.org/10.3109/14992021003753482

Brand, T., & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *Journal of the Acoustical Society of America*, *112*(4), 1597–1604. https://doi.org/10.1121/1.1502902

Brendel, A., Haubner, T., & Kellermann, W. (2020). A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis. *IEEE Transactions on Signal Processing*, *68*, 3545–3558. https://doi.org/10.1109/TSP.2020.3000199

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2009). Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers. *Journal of the Acoustical Society of America*, *125*(6), 4006–4022. https://doi.org/10.1121/1.3117686

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*, 975–979. https://doi.org/10.1121/1.1907229

Delcroix, M., Ochiai, T., Žmolíková, K., Kinoshita, K., Tawara, N., Nakatani, T., & Araki, S. (2020). Improving speaker discrimination of target speech extraction with time-domain speakerbeam. In *Proceeding of the international conference on acoustics, speech and signal processing (ICASSP)* (pp. 691–695). Barcelona, Spain. https://doi.org/10.1109/ICASSP40776.2020.9054683

Delcroix, M., Žmolíková, K., Ochiai, T., Kinoshita, K., & Nakatani, T. (2021). Speaker activity driven neural speech extraction. In *Proceeding of the international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6099–6103). Toronto, Canada. https://doi.org/10.1109/ICASSP39728.2021.9414998

Deng, C., Ma, S., Sha, Y., Zhang, Y., Zhang, H., Song, H., & Wang, F. (2021). Robust speaker extraction network based on iterative refined adaptation. In *Proceeding of the interspeech*

(pp. 3530–3534). Brno, Czechia. https://doi.org/10.21437/Interspeech.2021-2250

Dillon, H. (2012). *Hearing aids*. Thieme Medical Publishers.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., & Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, *37*(4), 1–11. https://doi.org/10.1145/3197517.3201357

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America*, *106*(6), 3578–3588. https://doi.org/10.1121/1.428211

Ge, M., Xu, C., Wang, L., Chng, E. S., Dang, J., & Li, H. (2020). SpEx+: A complete time domain speaker extraction network. In *Proceeding of the interspeech* (pp. 1406–1410). Shanghai, China. https://doi.org/10.21437/Interspeech.2020-1397

Gu, R., Chen, L., Zhang, S. X., Zheng, J., Xu, Y., Yu, M., Su, D., Zou, Y., & Yu, D. (2019). Neural spatial filter: Target speaker speech separation assisted with directional information. In *Proceeding of the interspeech* (pp. 4290–4294). Graz, Austria. https://doi.org/10.21437/Interspeech.2019-2266

Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *Proceeding of the International conference on acoustics, speech and signal processing (ICASSP)* (pp. 31–35). Shanghai, China. https://doi.org/10.1109/ICASSP.2016.7471631

Hochmuth, S., Kollmeier, B., & Shinn-Cunningham, B. (2018). The relation between acoustic-phonetic properties and speech intelligibility in noise across languages and talkers. In *Proceedings of meetings on acoustics DAGA 2018* (pp. 628–629). Munich, Germany. https://pub.dega-akustik.de/DAGA_2018/data/articles/000514.pdf

Hygge, S., Ronnberg, J., Larsby, B., & Arlinger, S. (1992). Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. *Journal of Speech, Language, and Hearing Research*, *35*(1), 208–215. https://doi.org/10.1044/jshr.3501.208

ITU-T (2001) Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862. Technical report, international telecommunications union (ITU-T) Recommendation.

Kidd, G., Mason, C. R., Best, V., Roverud, E., Swaminathan, J., Jennings, T., Clayton, K., & Steven Colburn, H. (2019). Determining the energetic and informational components of speech-on-speech masking in listeners with sensorineural hearing loss. *Journal of the Acoustical Society of America*, *145*(1), 440–457. https://doi.org/10.1121/1.5087555

Kidd Jr, G., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., & Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. *Journal of the Acoustical Society of America*, *140*(1), 132–144. https://doi.org/10.1121/1.4954748

Kollmeier, B., & Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *Journal of the Acoustical Society of America*, *102*(4), 2412–2421. https://doi.org/10.1121/1.419624

Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017). Development of an adaptive scaling method for subjective listening effort. *Journal of the Acoustical Society of America*, *141*(6), 4680–4693. https://doi.org/10.1121/1.4986938

Le Roux, J., Wisdom, S., Erdogan, H., & Hershey, J. R. (2019). SDR–half-baked or well done? In *Proceeding of the international conference on acoustics, speech and signal processing (ICASSP)* (pp. 626–630). Brighton, UK. https://doi.org/10.1109/ICASSP.2019.8683855

Liu, K., Du, Z., Wan, X., & Zhou, H. (2023). X-Sepformer: End-to-end speaker extraction network with explicit optimization on speaker confusion. In *Proceeding of the international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1–5). Rhodes Island, Greece. https://doi.org/10.1109/ICASSP49357.2023.10095609

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, *27*(8), 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167

Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2020). Estimation of model-based predictions, contrasts and means. https://github.com/easystats/modelbased. R package on Github.

Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., et al. (2024). Evolving deep neural networks. In *Artificial intelligence in the age of neural networks and brain computing* (pp. 269–287). Elsevier.

Ohlenforst, B., Zekveld, A. A., Jansma, E. P., Wang, Y., Naylor, G., Lorens, A., Lunner, T., & Kramer, S. E. (2017). Effects of hearing impairment and hearing aid amplification on listening effort: A systematic review. *Ear and Hearing*, *38*(3), 267–281. https://doi.org/10.1097/AUD.0000000000000396

Pandey, A., & Wang, D. (2018). A new framework for supervised speech enhancement in the time domain. In *Proceeding of the interspeech* (pp. 1136–1140). Hyderabad, India. https://doi.org/10.21437/Interspeech.2018-1223

Parizet, E. (2002). Paired comparison listening tests and circular error rates. *Acta Acustica United with Acustica*, *88*(4), 594–598. https://api.semanticscholar.org/CorpusID:17210790

Reinten, I., De Ronde-Brons, I., Houben, R., & Dreschler, W. (2021). Measuring the influence of noise reduction on listening effort in hearing-impaired listeners using response times to an arithmetic task in noise. *Trends in Hearing*, *25*, 2331216 5211014437. https://doi.org/10.1177/23312165211014437

Rennies, J., Best, V., Roverud, E., & Kidd Jr, G. (2019). Energetic and informational components of speech-on-speech masking in

binaural speech intelligibility and perceived listening effort. *Trends in Hearing*, *23*, 1–21. https://doi.org/10.1177/2331216519854597

Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *Journal of the Acoustical Society of America*, *136*(5), 2642–2653. https://doi.org/10.1121/1.4897398

Rhebergen, K. S., Versfeld, N. J., Dreschler, W., et al. (2005). Release from informational masking by time reversal of native and non-native interfering speech. *Journal of the Acoustical Society of America*, *118*(3), 1274–1277. https://doi.org/10.1121/1.2000751

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification*, *12*(4), 283–299. https://doi.org/10.1177/1084713808325306

Sinha, R., Rollwage, C., & Doclo, S. (2024). Variants of LSTM cells for single-channel speaker-conditioned target speaker extraction. *EURASIP Journal on Audio, Speech, and Music Processing*, *2024*(1), 1–13. https://doi.org/10.1186/s13636-024-00384-0

Sinha, R., Scherer, A. C., Doclo, S., Rollwage, C., & Rennies, J. (2023). Subjective performance evaluation of single-channel speaker-conditioned target speaker extraction algorithms for complex acoustic scenes. In *Proceeding of the ITG conference on speech communication* (pp. 1–5). Aachen, Germany. https://doi.org/10.30420/456164019

Sinha, R., Tammen, M., Rollwage, C., & Doclo, S. (2022). Speaker-conditioning single-channel target speaker extraction using conformer-based architectures. In *Proceeding of the international workshop on acoustic signal enhancement (IWAENC)* (pp. 1–5). Bamberg, Germany. https://doi.org/10.1109/IWAENC53105.2022.9914691

Smeds, K., Wolters, F., & Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of the American Academy of Audiology*, *26*(02), 183–196. https://doi.org/10.3766/jaaa.26.2.7

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transaction on Audio, Speech, and Language Processing*, *19*(7), 2125–2136. https://doi.org/10.1109/TASL.2011.2114881

Thoidis, I., & Goehring, T. (2024). Using deep learning to improve the intelligibility of a target speaker in noisy multi-talker environments for people with normal hearing and hearing loss. *Journal of the Acoustical Society of America*, *156*(1), 706–724. https://doi.org/10.1121/10.0028007

Vincent, E., Virtanen, T., & Gannot, S. (2018). *Audio source separation and speech enhancement*. NJ, USA: John Wiley & Sons.

Wagener, K., Brand, T., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III : Evaluation des Oldenburger Satztests (Development and evaluation of a German sentence test Part III : Evaluation of the Oldenburg sentence test). *Zeitschrift für Audiologie*, *38*(3), 86–95. https://archive.org/details/zf-a-1999-38-3-086-095-original

Wang, A. A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *Proceeding of the automatic speech recognition and understanding workshop (ASRU)* (pp. 577–582). Virgin Islands, USA. https://doi.org/10.1109/ASRU.2003.1318504

Wang, F. L., Lee, H. S., Tsao, Y., & Wang, H. M. (2022). Disentangling the impacts of language and channel variability on speech separation networks. In *Proceeding of the interspeech* (pp. 5343–5347). Incheon, Korea. https://doi.org/10.21437/Interspeech.2022-509

Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J. R., Saurous, R. A., Weiss, R. J., Jia, Y., & Moreno, I. L. (2019). VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Proceeding of the interspeech* (pp. 2728–2732). Graz, Austria. https://doi.org/10.21437/Interspeech.2019-1101

Wichern, G., Antognini, J., Flynn, M., Zhu, L. R., McQuinn, E., Crow, D., Manilow, E., & Roux, J. L. (2019). Wham!: Extending speech separation to noisy environments. In *Proceeding of the interspeech* (pp. 1368–1372). Graz, Austria. https://doi.org/10.21437/Interspeech.2019-2821

Xu, C., Rao, W., Chng, E. S., & Li, H. (2019). Time-domain speaker extraction network. In *Proceeding of the automatic speech recognition and understanding workshop (ASRU)* (pp. 327–334). Sentosa, Singapore. https://doi.org/10.1109/ASRU46091.2019.9004016

Xu, C., Rao, W., Chng, E. S., & Li, H. (2020). SpEx: Multi-scale time domain speaker extraction network. *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, *28*, 1370–1384. https://doi.org/10.1109/TASLP.2020.2987429

Zhang, W., Yang, L., & Qian, Y. (2023). Exploring time–frequency domain target speaker extraction for causal and non-causal processing. In *Proceeding of the automatic speech recognition and understanding workshop (ASRU)* (pp. 1–6). Taipei, Taiwan. https://doi.org/10.1109/ASRU57964.2023.10389752

Žmolíková, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., & Černocký, J. (2019). Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, *13*(4), 800–814. https://doi.org/10.1109/JSTSP.2019.2922820

Žmolíková, K., Delcroix, M., Ochiai, T., Kinoshita, K., Černockỳ, J., & Yu, D. (2023). Neural target speech extraction: An overview. *IEEE Signal Processing Magazine*, *40*(3), 8–29. https://doi.org/10.1109/MSP.2023.3240008