Real-time multichannel deep speech enhancement in hearing aids: Comparing monaural and binaural processing in complex acoustic scenarios

Nils L. Westhausen, Hendrik Kayser, Theresa Jansen, Bernd T. Meyer

***Abstract*—Deep learning has the potential to enhance speech signals and increase their intelligibility for users of hearing aids. Deep models suited for real-world application should feature a low computational complexity and low processing delay of only a few milliseconds. In this paper, we explore deep speech enhancement that matches these requirements and contrast monaural and binaural processing algorithms in two complex acoustic scenes. Both algorithms are evaluated with objective metrics and in experiments with hearing-impaired listeners performing a speech-in-noise test. Results are compared to two traditional enhancement strategies, i.e., adaptive differential microphone processing and binaural beamforming. While in diffuse noise, all algorithms perform similarly, the binaural deep learning approach performs best in the presence of spatial interferers. Through a post-analysis, this can be attributed to improvements at low SNRs and to precise spatial filtering.**

***Index Terms*—binaural, low-latency, multi-channel, real-time, speech-enhancement, subjective evaluation**

I. Introduction

The improvement of speech intelligibility for hearing-aid users is a challenging problem in complex acoustic scenes. Hearing aids often feature multiple microphones on each device, which enables multi-channel processing for the left and right side. Signals captured at the relatively densely positioned microphones (with a distance of approximately 1 cm) at one side can be enhanced, resulting in monaural processing; alternatively, true binaural processing can be performed, which requires a transmission of the microphone signals between the two devices.

One of the most widely adopted algorithms for monaural processing in this context is bilateral adaptive differential microphones (ADM) [1]. The ADM places the region(s) of the least sensitivity in the rear hemisphere of the listener and amplifies sound in the frontal hemisphere to improve the signal-to-noise ratio (SNR). The monaural processing of the ADM can result in slight distortion of binaural information [2]. A traditional approach for binaural processing is the binaural Minimum Variance Distortionless Response beamformer (MVDR) [3] using mocrophones on both sides of the head, with one instance of the MVDR beamformer running for each side based on the same parameters for directional steering and noise covariance. While the binaural MVDR is very capable to enhance the target direction, it distorts the binaural cues for sources not radiating from the target direction.

Recently, deep-learning based speech enhancement approaches have shown promising results in the context of hearing aids [4]–[6]. Several studies evaluated deep speech enhancement with hearing aids using headphones and audio material that was created offline [7]–[9]. Recent studies evaluated signals processed in real time [10], [11], which resulted in promising intelligibility improvements; but they were also limited to single-channel enhancement and did not use the signals of hearing aid microphones. A number of additional approaches were suggested for the Clarity Enhancement Challenges [9], [12]. The aim of the clarity challenges is to evaluate machine learning approaches for speech-in-noise conditions in the context of hearing aids. The first and the second enhancement round used simulated hearing aid signals, where in the first only one interferer (localized speech or localized noise) was present, while the second round featured multiple interfering sources. The systems that were successful in the first challenge were trained for more traditional speech enhancement with a target approximately from the front, while in the second challenge, because of the more complex scenario, the well-performing approaches were trained for target speaker extraction such as in [13]. While all proposed systems were causal, the challenges did not aim for efficient models that can be applied in real-time on restricted hardware.

In our earlier work, we proposed an approach that meets two important requirements of hearing devices, i.e., low-latency processing and limited computational resources. We introduced the binaural group communication filter-and-sum network (GCFSnet), which was applied to the problem of speaker *separation* [14]. The approach used grouping of latent representations and weight sharing between these groups for a decreased computational complexity while maintaining an algorithmic latency of 2 ms. The model had full access to the bilateral channels of hearing aids without latency and produced dichotic output to preserve spatial cues. However, a working implementation of such a bilateral communication link would require a wired connection between the hearing devices, which is not desirable.

To address this issue, we evaluated the GCFSnet approach in

Nils L. Westhausen and B. T. Meyer are with the Communication Acoustics group and the Cluster of Excellence "Hearing4all", Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany (e-mail: {nils.westhausen, bernd.meyer}@uni-oldenburg.de).

Hendrik Kayser is with the Auditory Signal Processing and Hearing Devices group and the Cluster of Excellence "Hearing4all", Carl von Ossietzky Universität Oldenburg, and the Hörzentrum Oldenburg gGmbH Oldenburg, Germany (e-mail: hendrik.kayser@uni-oldenburg.de).

Theresa Jansen is with the Hörzentrum Oldenburg gGmbH and the Cluster of Excellence "Hearing4all", Oldenburg, Germany (e-mail: jansen@hz-ol.de).

combination with a low-bitrate binaural link between bilateral models [15] and applied it to speech enhancement. The models were designed to be compatible with hearing-aid chips used for research (in terms of computational complexity) [16]. While these approaches resulted in promising improvements compared to models without any link and only small performance loss compared to having no delay in the transmission, the evaluations were limited to objective metrics and the experiments were performed with simulated test data from the same domain as the training data; it is therefore unclear how well the algorithms generalize to unseen conditions.

In the current study, we explore if speech enhancement with the GCFSnet is beneficial for hearing-impaired, aided listeners with moderate to severe hearing loss. To quantify the contribution of binaural communication between hearing aids, we compare two versions of the network, both of which produce dichotic output which is either based on binaural features (GCFSnet(b)) or on bilateral processing that only relies on monaural features (GCFSnet(m)). Two current strategies in hearing-aid processing are used as baseline: ADM and a fixed, i.e., non-adaptive in terms of noise statistics and direction, binaural MVDR beamformer. Additionally, subjective experiments cover an un-aided condition and only gain-adjusted signals individually for each subject. The subjective target metric is speech intelligibility quantified with a speech-in-noise test during a testing procedure that dynamically adapts the SNR [17], [18]. The goal of this test is to determine the SNR that yields 80% speech intelligibility, i.e., the speech reception threshold (SRT80).

All algorithms are evaluated in two complex acoustic scenes rendered in real time to 16 loudspeakers using an open-source software to simulate acoustic environments [19]. Listeners are supplied with hearing-aid dummies of the Portable Hearing Lab (PHL) [20], a portable and wearable research platform for hearing aid algorithms and signal processing. The dummies are controlled by th research open-source hearing aid software open Master Hearing Aid (openMHA) [21] running on a standard PC. With this setup, effects of individual heads are taken into account. The first scene includes a target speaker in front and two interfering speakers from $\pm 60°$ azimuth. The second scene also covers a target speaker in front but uses diffuse cafeteria noise.

We are also interested in the predictive power of objective metrics relative to subjective measurements and therefore perform an evaluation with two measures, the hearing aid speech perception index (HASPI w2) [22] and the modified binaural short time objective intelligibility [23]. The listeners' hearing loss is considered by HASPI by using subject-specific audiogram data as additional input; similarly, the Cambridge MSBG hearing loss model [24] is combined with MBSTOI [23] to obtain individual predictions.

Finally, the GCFSnet models are trained on large amount of synthetic static scenes with a simulation based on the measured head related transfer function (HRTF) set of the same hearing aid dummies on an artificial head (without any movement simulation). The subjective measurements described above should be informative if algorithms trained in these specific conditions could generalize to scenes that include head movements or effects from individual physical traits such as head geometries.

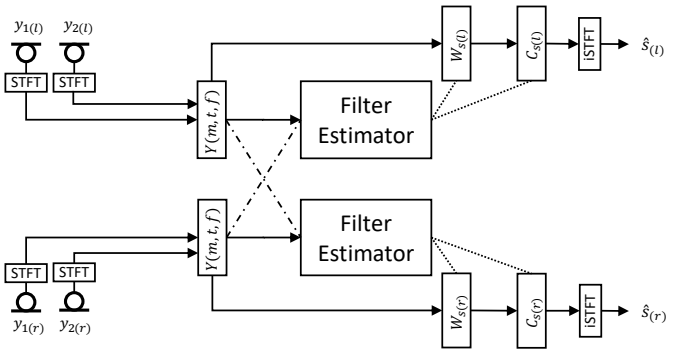

Fig. 1. Illustration of the proposed approach for spatial filtering and post filtering. The filters for the left and right side are estimated by separate models. The dashed-dotted line symbolizes an optional exchange of the complex TF representation of microphone signals for binaural input features.

II. Methods

A. Deep spatial and post filtering

In this study, we focus on scenarios with one target speaker in front and additional maskers, which can be either diffuse noise or two interfering speakers. Our enhancement approach uses the STFT representation of the mixture y as input, which can be written as

$$Y(m,t,f) = X_S(m,t,f) + X_D(m,t,f). \tag{1}$$

In this notation, Y denotes the STFT of the mixture where as m, t, f are the microphone, frame and frequency index, respectively. X_S is the reverberant STFT of the target speech s convolved with the impulse response of the target speaker h_s. X_D corresponds to the STFT of the interfering sources d.

We aim to extract the source estimate $\hat{s}$ including the direct part as well as early reflections. The extraction is performed for the left and right side separately with complex filter-and-sum beamforming

$$\tilde{S}(t,f) = \sum_{m=1}^{M} Y(m,t,f) \cdot W(m,t,f). \tag{2}$$

$\tilde{S}$ denotes an intermediate time-frequency (TF) representation of the estimated target speaker and $W \in \{z \in \mathbb{C} \mid -r \leq \Re(z) \leq r, -r \leq \Im(z) \leq r\}$ corresponds to the complex filter weights estimated by the model where r is the learned range of the real and imaginary components. M is the number of microphones used for the filtering. In the next step, a single frame postfilter is applied with the aim to further improve the signal

$$\hat{S}(t,f) = \tilde{S}(t,f) \cdot C(t,f). \tag{3}$$

$C \in \{z \in \mathbb{C} \mid -r \leq \Re(z) \leq r, -r \leq \Im(z) \leq r\}$ are the complex filter weights, where the numeric range of filter coefficients r is a learnable parameter. $\hat{S}$ is transformed back to the time domain by an iSTFT. The general structure of the filtering is illustrated in Figure 1.

B. Architecture

In this study, speech enhancement is performed with the *group communication filter and sum network* (GCFSnet) introduced in [15]. The architecture is illustrated in Figure 2; it contains five modules and uses group communication [25] and weight sharing for a reduced computational footprint. First, the feature vector of size B is scaled by a single learned parameter in the grouping module. This feature vector is then projected by a fully connected (FC) layer with shape $B \times P$ and a tanh activation function. The latent representation of size P is split in G equal groups.

The second step is represented by the conv module which is shared between groups. In this module, an FC layer with tanh activation and dimension $(P/G) \times U$ maps the latent representation of the group to the hidden size U. The representation of size U is processed by two causal depth-wise separable convolution layers (DS-Conv), the first with kernel size 5 and the second with kernel size 3 in time dimension. Inspired by [26], a skip connection with a depthwise convolution (D-Conv) with kernel size 1 is implemented in parallel to the DS-convs. The idea of the conv module is to capture information on a shorter time scale.

The next step is a group communication (GC) block. In our own previous work ([15] and [14]) GC was implemented with transform average concatenate (TAC) as suggested in [27]. For the current study, we exchanged TAC with a simpler structure we refer to as group communication with group mixing: First, a FC layer of shape $U \times (P/G)$ with tanh activation maps the hidden representations of the groups to size (P/G). The mapped representations are concatenated to form a latent representation of size P. This representation is mixed by an $P \times P$ FC layer with tanh activation. Next, this intermediate representation is split in G groups and mapped by an FC layer with shape $(P/G) \times U$ and tanh activation. In the final step of this block, the input to the mixing module is added to the output as a skip connection. This group mixing mechanism is slightly less complex compared to the TAC with respect to implementation and is restricted to fixed-range intermediate representations (due to the use of the tanh activation), which can be advantageous for fixed-point implementations.

The third module is the shared GRU (gated recurrent units) module. This module includes two consecutive GRU layers with U units and an additive skip connection with a D-Conv for scaling. This module is intended to capture more long-term temporal information.

GC is again applied after the GRU module. Next, the hidden representations of the groups are mapped to size P/G by an FC layer of size $U \times (P/G)$. The groups are combined to form a latent representation of size P. In the final step of this module, the real and imaginary parts of the filter weights W and C are predicted by FC layers with tanh activations and scaled by the learned parameter r as mentioned in subsection II-A.

C. Training data generation

Robust speech enhancement models require extensive data for training. In this study, we generated fixed training and development sets. The training set comprised 160k samples, while the development set included 3k samples, each 4 seconds long with a 16 kHz sampling frequency. Speech data was sourced from the DNS-Challenge dataset [28], and noise data from three datasets: DNS-Challenge [28], the first Clarity enhancement challenge dataset [9], and LibriMix [29]. We also simulated 60k multichannel binaural room impulse responses (BRIR) using RAZR [30]. To enable the simulation of hearing-aid microphone signals for the hearing aid dummies of the PHL we measured an HRTF set with an angular resolution of 1° using a KEMAR dummy head. The setup for this measurement is described in [31]. The limitation of the loudspeakers in this setup made a low-frequency extension necessary which is explained in [32]. This set features the four hearing-aid microphones with the devices placed on the KEMAR artificial head, front-left, front-right, left-back and right-back.

Our simulated rooms varied in size, with widths ranging from 3 to 10 m, heights between 2.5 and 4 m, and surface areas from 12 to 100 m^2. The reverberation time (T60) ranged from 0.25 to 1 s. The receiver was positioned no more than one meter from the center of the room at a height of 1 to 1.4 m. Each scene featured a target speaker placed within an azimuth range of -10° to 10° azimuth. Two interfering speakers were placed at different angles while omitting the -20° to 20° azimuth range. A minimum angular separation of 10° between the speakers was enforced. The speakers' distance ranged from 0.75 to 2m, at heights of 1 to 1.4 m. Additionally, two noise sources were simulated: one localized interferer from the DNS-Challenge dataset or the Clarity dataset, and another for diffuse-like noise, created using the RAZR feedback delay network output combined with ambient noise from LibriMix, which is based on the WHAM! corpus [33]. Both noise sources were positioned at least 1 m away from the receiver, at heights of 1 to 1.4 m.

For each interfering speaker, a random better-ear SNR between -8 and 8 dB was sampled from an equal distribution. The noise sources were mixed at a relative SNR sampled from $\mathcal{N}(0, 5^2)$ dB. The full noise signal is then mixed with target speech with a better ear SNR equally distributed between -8 and 8 dB. The mix is scaled to a value drawn from $\mathcal{N}(-28, 10^2)$ dB relative to full scale. One part of the training and development set contained only two interfering speakers (30% of the scenes), another part contained only interfering noise (30%), and another part contained both speech and noise interferers (40%).

For the training target the BRIR of the front microphones for the target speaker was windowed to include the direct part of the BRIR as well as the some early reflections up to 300 ms. BRIRs are windowed as described in [34]. To obtain the final binaural target the target speech signal was convolved with the windowed BRIR.

D. Model and training configuration

Two versions of the GCFSnet were trained for this study, one with features only from the ipsilateral side (monaural version) and one where the model also received the features of the contralateral side (binaural version). The concatenated real

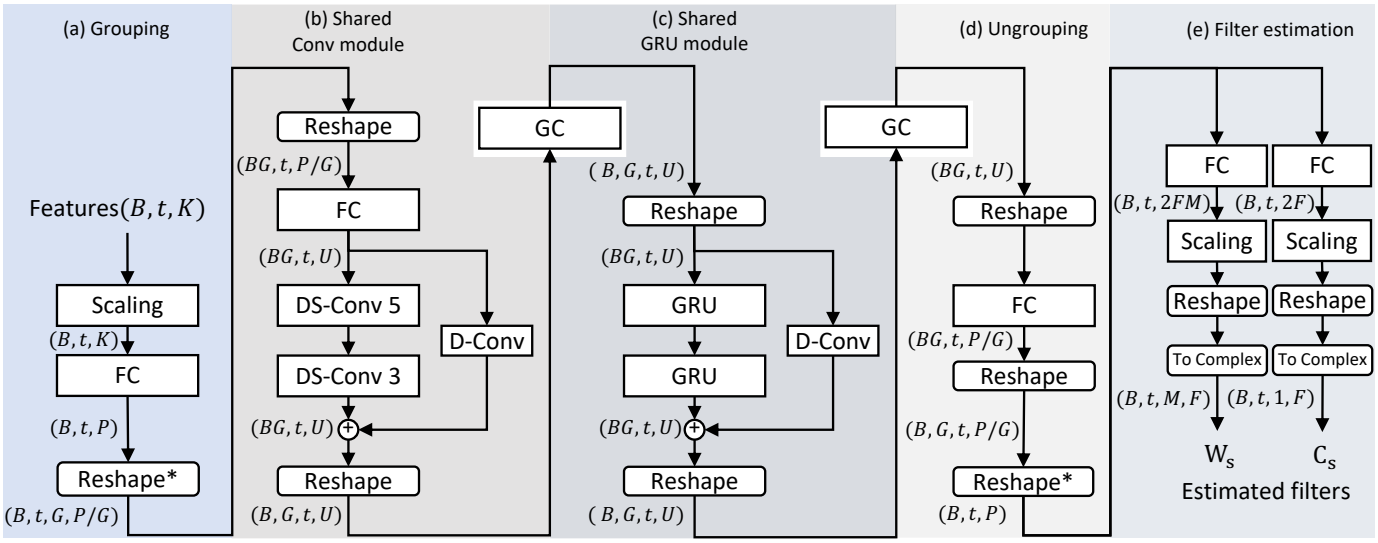

Fig. 2. Illustration of the proposed filter estimation model. Reshape operations with (*) include axes permutation.

and imaginary part of the frequency representation served as input features as suggested in [14]. For the binaural version, we assumed a transmission without delay between the ears. The models on both sides shared the same weights. Following [15], P was set to 128 and G was set to 8, and U was set to 32. This resulted in 135k weights for the monaural version and in 168k weights for the binaural version. The model sizes were chosen based on the constraints of the research hearing aid chip described in [16]. The models were trained with quantization-aware training [35] for weights and biases. The weights were linearly quantized to 8 bits and the biases to 16 bits with both having a range between -1 and 1. This reduced the size of the model drastically compared to keeping all weights and biases at 32 bits. The calculations inside the model were not quantized. The frame length was 4 ms and the frame shift was 2 ms. The FFT size was 128 with equal front and back padding. A hann window was applied for the forward transformation. These parameters were given by the signal processing chain of the openMHA setup used in this study.

The model was trained with compressed spectral Mean Squarred Error (cMSE) as proposed in [36]. The STFT configuration of the loss function is independent of the one used for the filtering framework since it was calculated after signal reconstruction. The STFT of the loss function was set a to window length of 20 ms with a 10 ms shift and an FFT size of 320. The model was trained for 50 epochs with an initial learning rate of 2e-3 and the ADAM optimizer in Tensorflow 2.11. In each epoch, the learning rate was multiplied by 0.98. If the loss on the development set did not decrease for 5 consecutive epochs, the loss was multiplied by 0.5. For gradient clipping, AutoClip [37] was applied with $p = 10$ for smoother training and better generalization. The weights were saved when the loss on the development set decreased. For running the models inside the openMHA, the full architecture was implemented in standard lib C without any special optimizations.

E. Baseline algorithms

The two configurations of the GCFSnet were compared with two traditional algorithms commonly used in hearing aids. The first was the adaptive differential microphone (ADM) [1] which uses the two microphones in one hearing aid to create a monaural filter. The ADM was configured to optimize its output for placing the region of the least sensitivity (Null) in the rear hemisphere. ADM algorithms on the left and right side work independently, which can slightly distort the binaural perception.

The second was a fixed binaural Minimum Variance Distortionless Response beamformer (MVDR) [3] steered to 0° azimuth. This beamformer uses all four microphones available in the hearing aid setup without delay for the transmission. It aims at minimizing the power of the output signal with the constraint that a signal arriving from the front is preserved under the assumption of an isotropic, spatially diffuse noise field. It produces an output signal that only contains binaural cues related to its target direction, i.e., binaural cues related to sound sources located at other directions are not preserved.

F. Complex acoustic scenes

The measurements were conducted in two complex acoustic scenes rendered to a ring of 16 GENELEC 8030 loudspeakers with a radius of 1.56 m and a height of each loudspeaker of 1.18 m. The azimuth angle between the loudspeakers was 22.5°. The loudspeakers were connected to a RME ADI-8 Pro soundcard with a 44.1 kHz sampling frequency. The room had the dimensions of 5.93 m × 5.01 m × 2.74 m (length × width × height) and a low reverberation time of around 170 ms. The simulation was performed with TASCAR [19]. The scenes were rendered at a sampling frequency of 44.1 kHz. The first scene contained a target speaker from 0° azimuth and two localized interfering speakers from +60° and -60° azimuth (*S0N±60 IFFM*). Utterances from the Oldenburg sentence test with a female talker were used as target sentences, while the interfering signals were were based on the International

Female Fluctuating Masker (IFFM), a version of the International Speech Test Signal (ISTS) with shortened gaps [38]. The ISTS is an non-intelligible speech signal built from segments of Arabic, English, Mandarin, Spanish, German and French, which enables reproducible measurements. Interfering signals were continuously presented during each measurement trial at 65 dB SPL (sound pressure level). The level of the target was varied during the measurement to adjust the SNR. The second scene contained a target speaker from 0° azimuth and diffuse cafeteria noise (*S0Ndiff Cafeteria*). The cafeteria noise was taken from [39]. The noise is a first-order ambisonics recording performed at lunch time in the cafeteria at the University of Oldenburg. All intelligible speech segments were removed in post processing. As for the first scene, the continuous noise was presented at 65 dB SPL while the target speech level was varied.

G. Hearing-aid configuration

Hearing-aid processing was simulated with the open Master Hearing Aid (openMHA) [21] software running on a laptop CPU. Signals were recorded with hearing-aid dummies, each with two MEMS microphones, of the portable hearing lab (PHL) [20] worn by the listeners. A MOTU A8 sound card was used to capture the hearing aid input signals at a sampling frequency of 48 kHz. The microphones were calibrated with diffuse speech-shaped noise.

For signal enhancement, the recorded signals were downsampled to 16 kHz in the openMHA. The bilateral ADM processed the signals in the time domain, while the MVDR beamformer and the GCFSnet configurations operate in the frequency domain. The frame length of the hearing-aid processing was 2 ms, the window length was 4 ms and the FFT length was 8 ms corresponding to 128 samples. The enhancement algorithms can be bypassed for the condition *unprocessed*. Next, a multi-band dynamic range compressor (MBDRC) was applied for gain control and compression. The MBDRC was followed by a frequency shifter for feedback reduction. Next, an equalization was performed to compensate spectral effects caused by the closed coupling of the hearing aid receivers in the ear canal. Finally, the signal was resampled to 48 kHz and presented over the hearing-aid receivers.

Personal ear molds were manufactured for each subject for a closed hearing-aid fitting with receivers in the ear canal. The closed fitting reduces feedback and comb-filter effects and allows a higher gain in the low frequencies compared to an open fitting. Subjects were fitted with a loudness-based gain prescription rule [40]. The maximum allowed gain was set to 30 dB and the maximum output was limited at 100 dB. The measured algorithmic latency of the hearing aid setup was 5.4 ms and the latency of the audio hardware and server was 8.7 ms. This results in a total latency of 14.1 ms for the whole hearing aid setup.

H. Subjective measurement procedure

For evaluating the intelligibility of the enhancement algorithms, speech reception thresholds (SRT) were measured with the Oldenburg matrix sentence test (OLSA). The OLSA

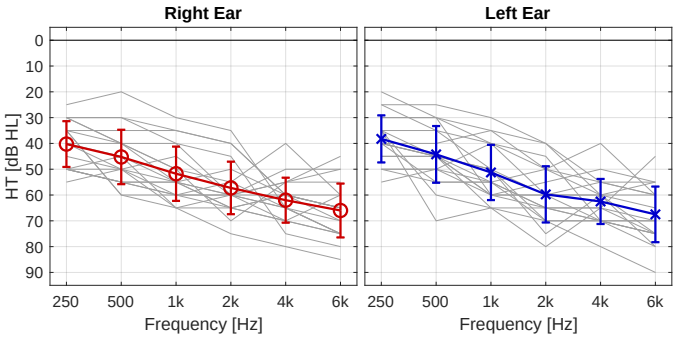

Fig. 3. Hearing thresholds (HT) in dB hearing level (HL) of all subjects. The colored lines show the mean HT and the standard deviation. Individual audiograms are shown in gray.

is a procedure where the SNR is changed adaptively during the measurement based on the subject's responses. The SRT is calculated by a least-squares fit. The speech material is structured in lists of 20 sentences where each list is used for one measurement trial. All sentences have the same structure which is [Name], [Verb], [Numeral], [Adjective], [Noun], with 10 alternatives per category. The subjects used a touch screen with the 5x10 alternatives to log their response.

In this study, the adaptation procedure is configured to measure the SRT80, i.e., the SNR with 80% of the words being correctly identified which results in SNRs more similar to SNRs encountered in real-world scenarios as when measuring the traditional 50% correct point.

The measurements were performed for 6 different conditions, i.e., without any signal modifications (*Unaided*), with prescribed gain (*gain only*), and the four enhancement algorithms (*ADM*, *MVDR*, *GCFSnet(m)*, *GCFSnet(b)*) each with the prescribed gain. For each subject, the SRT80 was measured twice for each condition, resulting in test and retest scores.

I. Subjects

The evaluation of the algorithms was performed with 20 hearing-impaired subjects (age: 49-81 years, 12 female listeners). All subjects had a symmetric hearing loss corresponding to N3 or N4 after the Bisgaard standard audiograms [41]. N3 corresponds to sloping, moderate hearing loss and N4 to a sloping, moderate to severe hearing loss. All subjects were native German speakers and were fitted with hearing aids for more than 24 months. They gave their written consent prior to inclusion in the study. The experiment was approved by the ethics committee ("Kommission für Forschungsfolgenabschätzung und Ethik") of the Carl von Ossietzky Universität in Oldenburg, Germany (Drs.EK/2021/031-05). The mean hearing loss and standard deviation over all subjects is shown in Figure 3.

J. Objective evaluation

In addition to the subjective evaluation with hearing-impaired subjects, an objective evaluation was performed based on two metrics. The first is the better ear hearing aid speech perception index with a mapping function related to a

keyword recognition task (HASPI w2) [22]. The second is a combination of MSBG hearing loss model [24] and the modified binaural short time objective intelligibility (MBSTOI) [23] as previously used in the first round of the first Clarity enhancement challenge [9].

These two metrics were chosen since they relate to speech intelligibility prediction and allow the consideration of individual hearing loss. HASPI was specifically built to evaluate hearing aid algorithms, while MBSTOI explicitly takes binaural perception into account.

Both metrics require clean reference signals without the additive noise and simulated reverberation. These signals are not part of the subjective measurements and were therefore simulated. To this end, a binaural synthesis was conducted using TASCAR to simulate the acoustic scenes. This synthesis was performed using HRTFs of the hearing aid dummies on a KEMAR artificial head measured inside the loudspeaker ring used during the subjective measurements. The simulation enabled us to create separate, time-aligned reference signals for the objective metrics. To obtain the reference signal, we deactivated the reflecting surfaces, the image source model and the feedback delay network of TASCAR, so the reference contains no reverberation.

We simulated the signals at fixed SNRs. The metrics were averaged between -5 and 10 dB SNR (with 1 dB step) since this is an SNR range that covers challenging listening environments in which hearing-aid users would most likely profit from improvements. At each SNR, 20 OLSA sentences were synthesized at 44.1kHz. Each sentence had a random noise sample drawn from the original noise signals. The mixed signals were resampled to 48 kHz to match the input sampling frequency of the hearing aid processing. All simulated signals were processed with the openMHA setup individually for each subject including hearing loss compensation and enhancement conditions but with deactivated equalization for the hearing aid receivers since the signals are not played back. The results are averaged over all 20 sentences for each SNR.

Additionally, we evaluated the attenuation of the enhancement algorithms depending on the incidence angle. This evaluation was based on the TASCAR setup for the objective evaluations described above. We rendered 20 OLSA sentences for incidence angles from -180 to 180 in 5 degree steps. The source had a level equivalent to 70 dB SPL. The simulated room was not changed besides deactivating the noise sources so that the effects of room reverberation are taken into account. The attenuation was calculated as the ratio between the energy of the signals processed by the enhancement algorithms and signals processed without activated algorithms. The result was averaged over all 20 sentences.

III. Results

Figure 4 shows violin plots of the results of the subjective SRT measurements for all subjects, conditions and scenes together with the objective evaluation for the same subjects performed with HASPI w2 and MSBG+MBSTOI.

Results of subjective measurement of speech intelligibility: The statistical analysis of the listening experiments showed that all results are normally distributed according to a Kolmogorov-Smirnov test ($p \leq 0.05$). A one-way ANOVA was performed using subjective SRT scores (test and retest) for the different processing conditions, i.e., the condition *Unaided* was excluded. For the first acoustic scene with spatial maskers (IFFM), the ANOVA indicated a significant difference of algorithm-specific medians ($p \leq 0.05$). From the two best-performing algorithms (*GCFSnet(m)*) and *GCFSnet(b)*), *GCFSnet(b)* performs significantly better (with a median SRT80 -0.7) according to a paired t-test ($p \leq 0.05$) and achieves a median SRT that is 2.2 dB lower (better) than *GCFSnet(m)* and 2.6 dB lower compared to the best traditional algorithm (*ADM*). The SRT improvement of *GCFSnet(b)* over *Unaided* and *Gain only* is 9.0 dB and 5.3 dB, respectively.

In the diffuse noise scenario, a one-way also ANOVA indicated significant differences between the algorithm-specific medians ($p \leq 0.001$). However, when only including the four best-performing algorithms, differences were not significant. Among these, *MVDR* achieved the most favorable SRT at 1.6 dB, followed by *GCFSnet(b)* with an SRT of 1.8 dB, and *GCFSnet(m)* with 2.0 dB. *ADM* produces a slightly higher SRT of 2.5 dB. Notably, all enhancement algorithms demonstrated improvements in comparison to the *Unaided* condition, which had an SRT of 8.1 dB, and the *Gain only* condition, with an SRT of 6.3 dB.

Results in terms of HASPI: A similar statistical analysis was performed for the HASPI scores: A one-way ANOVA indicated a significant difference among the algorithms in both acoustic scenes ($p \leq 0.001$ in both cases). In the *S0N±60 IFFM* scene, similar to the subjective results, *GCFSnet(b)* achieved the best performance with the highest median HASPI score of 0.66. Its performance was significantly better than the second-best algorithm *MVDR* with a HASPI score of 0.5 (based on a paired t-test, $p \leq 0.001$). For the scene *S0Ndiff Cafeteria*, the performance differences between *MVDR*, *GCFSnet(m)*, and *GCFSnet(b)* did not reach a statistical significance. Among these, *GCFSnet(m)* and *GCFSnet(b)* recorded the highest HASPI scores of 0.48, followed by *MVDR* and *ADM* with scores of 0.44 and 0.33, respectively.

Overall Results in terms of MBSTOI: The MBSTOI results are shown in column (c) of Figure 4. We performed the same statistical analysis as described above, which indicated significant differences between processing strategies ($p \leq 0.001$). In both acoustic scenes, *GCFSnet(b)* achieved the highest median (0.49 and 0.44 for the IFFM and Cafeteria scene, respectively) which in both cases was significantly better than the runner-up *GCFSnet(m)* (paired t-test, $p \leq 0.001$). For the Cafeteria scene, this is different from the subjective results, where MVDR performed best and was not statistically different from the best-performing algorithms.

Correlation between subjective and objective results: To quantify how well objective metrics can predict responses from aided, hearing-impaired listeners, we analyze the Pearson correlation between the SRT80 from listening experiments and the according score obtained from HASPI and MBSTOI. We report the correlation r_{sub} that takes into account data from individual subjects in all conditions (i.e., all colored data points in Figure 4) as well as the correlation r_{med}

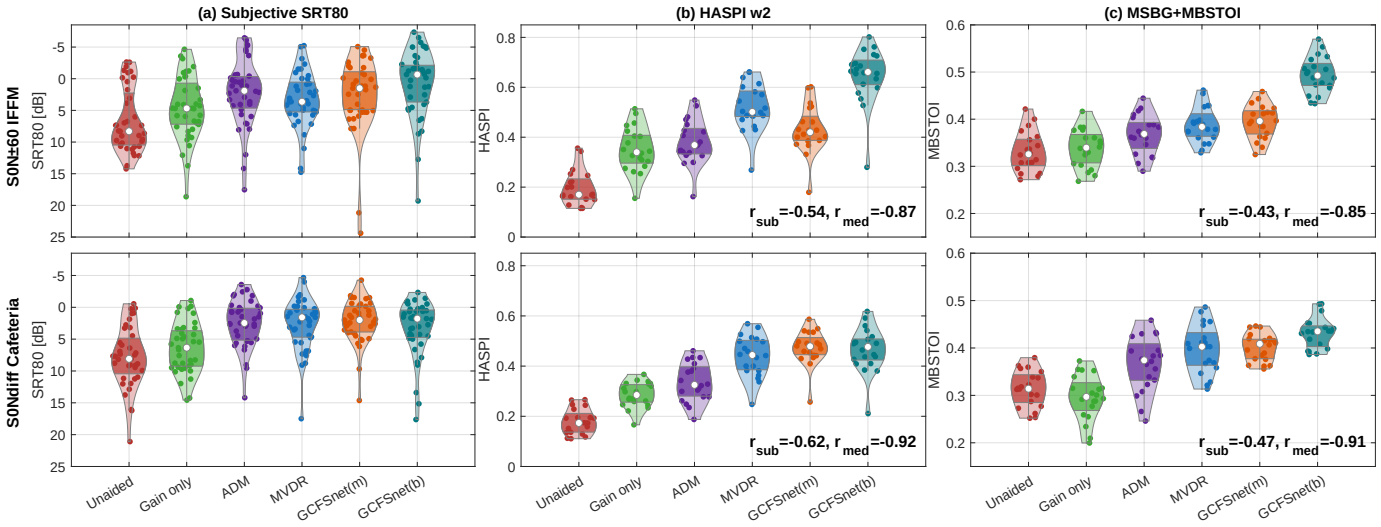

Fig. 4. Results for different HA enhancement strategies for subjective listening tests (left column) and objective metrics HASPI (middle column) and MBSTOI (right column). The axes are reversed for the left column. Violin plots are shown for two acoustic scenes (top and bottom row, respectively). The r-values show the correlation of the objective metric with the subjective measurements, both on subject level (r_{sub}) and for the medians (r_{med}).

based on median values in which the subject-level scores are pooled. The corresponding scores are shown in the lower right corners of the HASPI and MBSTOI panels in columns (b) and (c). All correlation values are significant at $p \leq 0.05$. Negative correlation coefficients are obtained since a better subjective performance results in a *lower* SRT80 value; note that the scales in the left column of Figure 5 are inverted. The correlations for individual subjects are moderate, while very strong correlations are obtained when the medians are considered.

Results metrics depending on SNR: In a post analysis, we explored the objective metrics depending on the SNR to gain insight how the enhancement algorithms perform in different noise conditions. Figure 5 shows the HASPI and MBSTOI scores plotted against the SNR for each scene individually. The results are pooled over all listeners, which are considered through their individual hearing loss that is used as input to the metrics. For *S0N±60 IFFM*, *GCFSnet(b)* produces a benefit in an SNR range of -5 to 5, for both HASPI and MBSTOI. For high SNRs, the metrics saturate and differences become very small. At low SNRs, *GCFSnet(m)* is predicted as second-best algorithm, but is outperformed at higher SNRs. The *MVDR* shows a steeper performance increase with SNR than *ADM* or *GCFSnet(m)* in both metrics. In the *S0Ndiff Cafeteria* conditions, the objective scores for *GCFSnet(b)*, *GCFSnet(m)* and the *MVDR* are very similar, i.e., both HASPI and MBSTOI do not predict SNR-specific gains in this diffuse-noise condition. *ADM* and *Gain only* produce lower objective scores, with a slight benefit of *ADM* over *Gain only* at low SNRs.

Results for attenuation over incidence angle: Since all algorithms perform multi-channel processing and spatial filtering, we analyze the attenuation depending on the azimuth angle. The results are shown in Figure 6.

The attenuation patterns of the algorithms are quite different: *GCFSnet(b)* exhibits the narrowest beam combined with

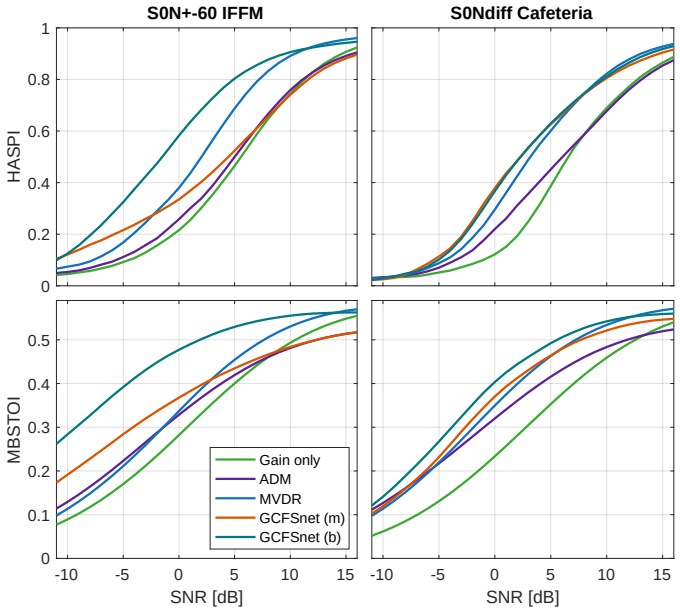

Fig. 5. Objective metrics in terms of HASPI (first rown) and MBSTOI (second row) pooled over all subjects plotted against SNR for both acoustic scenes.

strong attenuation: The full attenuation of -30 dB is already reached at ±45°. *GCFSnet(m)* also produces strong attenuation for angles deviating from 0° but with a clearly broader beam. An angle-dependent attenuation is also observed for both *ADM* and *MVDR*, but the overall attenuation is lower (with a lowest value of -13 dB). The beam patterns are slightly asymmetric, which could results from the measured HRTFs or the simulation of the room.

IV. Discussion

Comparison of performance of enhancement strategies: For the *S0N±60 IFFM* scene, the subjective and the objective metrics show a significant benefit for *GCFSnet(b)*, suggesting

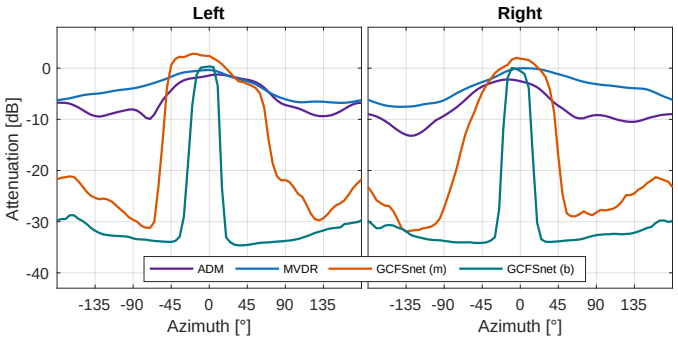

Fig. 6. Attenuation of clean speech over incidence angle for the left and the right hearing aid device separately.

that the binaural deep learning approach is a good choice in the presence of localized speech interferers. A possible contribution to this result is the narrow beam associated with *GCFSnet(b)* as depicted in Figure 6. The maximum attenuation is already reached at the location of the interferers at $\pm 60°$. It seems that the binaural features enable the model to accurately identify the direction of a speech source and to attenuate other sources outside the azimuth range defined during the training. Similarly, Figure 5 shows a clear benefit over large SNR range for *GCFSnet(b)*. In contrast, *GCFSnet(m)* does not produce benefits compared to the traditional methods for the *S0N±60 IFFM* scene. A potential reason becomes apparent in Figure 6: The monaural model produces an attenuation pattern wider than its binaural counterpart and does not reach the same level of attenuation, i.e., the suppression of interfering speakers will be less effective. However, the monaural approach still reaches attenuation levels of 20 dB for clean speech or more although it is limited to the signals of two closely-spaced microphones. In the scene with diffuse noise, we did not observe significant differences between traditional and neural systems, except for the gain-only condition, which was outperformed by the other algorithms. It seems plausible that the narrow beam of *GCFSnet(b)* does not provide strong benefits in conditions with diffuse noise only.

ADM produces good subjective results in both acoustic scenes, which could be linked to the partial preservation of binaural cues and a processing strategy that does not introduce stronger speech distortions. The performance of ADM is underestimated by the objective measures, especially for HASPI. This could be linked to the fact that HASPI does not consider binaural effects apart from better-ear listening and both metrics could not consider learning effects. Several subjects reported that the deep-learning based processing sounded unusual, which could indicate that GCFSnet algorithms produce artifacts different from ADM and MVDR. We presume that a longer exposure to GCFSnet processing could help listeners to adapt to these artifacts, potentially increasing performance with these algorithms.

Relation between objective and subjective results: The results of the objective evaluation of SRT80 and the evaluation in terms of HASPI and MBSTOI following a similar pattern as shown in Figure 4. On the individual level, only moderate correlations are observed (with r ranging from -0.43 to -0.62). We assume this can be partially attributed to variability caused by factors related to hearing impairment beyond the audiogram, or fatigue and the cognitive condition, which are not accessible to the models. When this variability is partially removed by calculating the median performance, very strong correlations of 0.85 and higher are obtained, which suggests that the objective metrics are a good indicator for the enhancement benefit when data is pooled over subjects. We find this encouraging since the objective measures used a binaural, synthetic input generated with TASCAR at a fixed SNR set, and head differences or head movements have not been considered.

Generalization between training, measurement and simulation setup: Some properties of the training data were similar to the testing condition (e.g., the subset with interfering speakers always contained two speakers) while most parameters differed (by using random random rooms, SNRs, source and receiver positions, including their height) and others were chosen to be disjunct from the test set (speaker identity and consequently signals, interferer signals). The training data only contained static scenes based on BRIR simulations combined with a single HRTF set of the hearing-aid dummies measured on an artificial head. On the other hand, the subjective evaluation covered listeners with individual head geometries, natural head movements and a real-time hearing aid signal processing pipeline. Given these differences, it was unclear if the algorithm could perform well in the test condition. The results suggest that the GCFSnet approach is able to generalize from the training to the measurements, especially for the *S0N±60 IFFM* scene with localized interfering speakers. It seems that head geometries and head movements are not a large issue despite being trained on static artificial scenes with only one head. For *S0Ndiff Cafeteria*, the models generalize well enough to achieve a similar performanceas the traditional methods. The models seems to also generalize to the simulation setup of the objective evaluation, which shows that the model can also generalize to another HRTF set (also measured on KEMAR but with a completely different loudspeaker and audio setup) and another simulation tool. It is unclear if the relatively small model size plays an important role in this context: For a small model, it might be difficult to handle complex acoustic conditions; on the other hand, smaller models are less prone to overfitting. Hence, it remains to be seen if better results can be obtained by changing the size of the model, or exposing it to different HRTFs and dynamic acoustic scenes during training.

Limitations and future work: The proposed binaural GCFSnet meets several requirements of real-world applications that were also reflected in the subjective measurements, such as providing benefits in free-field conditions and complex acoustic environments, while processing signals in real time. However, we also used a setup that enabled access to binaural signals without any latency, which is not compatible with wireless connections between hearing aids, something that is preferred over a wired connection by most hearing aid users. It is unclear how latency and lossy signal transmission would affect GCFSnet(b) and the MVDR beamformer. In our own related work, we have shown that the monaural

GCFSnet can profit from a low-bitrate binaural link with delayed binaural features [15] and hence we assume that a neural system with low-bitrate binaural communication would exhibit a performance similar to GCFSnet(b). It will be very interesting to quantify the effect of such a link in future research.

The monaural GCFSnet does not demonstrate a significant advantage over the traditional bilateral ADM, despite its higher overall complexity. The overall performance might be improved when optimizing the training set, something that was not done in our study. Although the training data covered challenging SNRs coupled with a wide array of reverberation times, training with dynamic scenes might improve the model. Further, given the high variability of subjective scores, we assume that individual listeners could profit more from GCFSnet than from ADM (and the other way round), so the proposed algorithm could be added as an alternative hearing program in the future.

Another limitation is still relatively high latency of the complete setup which is mainly resulting from the traditional audio hardware used during the evaluation. This latency could be reduced by porting the algorithms directly to the PHL platform. Since the current PHL has only very limited resources available it was not possible to perform the required amount of optimization for the GCFSnet algorithms in the scope of this study. The computational cost could be lowered by implementing GCFSnet using fixed-point operations only, which could also result in an implementation that is fully compatible with a research hearing aid chip such as the smartHeap [16].

The configurations of GCFSnet evaluated in this study were trained with a fixed steering towards the front, which requires the hearing aid user to directly face the target speaker. While this approach effectively addresses the issue of locating the attended source, it imposes a somewhat unnatural behavioral constraint on the subjects. Ideally, a more adaptable steering mechanism for GCFSnet would be beneficial as suggested in [42] and [43]. However, with increased flexibility comes the challenge of selecting the speaker that should be attended. Such a selection could be made with an external device, such as a smartphone. Alternatively, more sophisticated methods involving the analysis of eye gazing [44] or brain activity (invasive [45] or non-invasive [46]) could be explored. Each of these approaches, though complex, has potential for enhancing the user experience and the functionality of GCFSnet in more dynamic and natural listening environments.

The scope of the current setup is confined to two specific types of interferers: competing speech and diffuse cafeteria noise, each featured in a distinct scene. In real-world scenarios, individuals are likely to encounter a combination of various interferers, rather than isolated types. The rationale behind the study design was to maintain control over the experimental settings and to systematically assess performance in these two distinctly different interference situations. However, for a more comprehensive approximation to real-life conditions, future studies should consider evaluating scenarios that incorporate a combination of interferers.

V. Conclusion

This study explored the potential of deep neural networks for speech enhancement in hearing aids, based on a recurrent network with low latency and a small computational footprint, the GCFSnet.

Speech intelligibility for hearing-impaired listeners is clearly improved with the GCFSnet compared to using unprocessed signals.

In the presence of localized interferers, deep enhancement based on binaural communication produces the highest intelligibility compared to established hearing aid signal enhancement strategies considered here.

For data pooled over HI listeners, the objective metrics HASPI and MSBG+MBSTOI are good predictors for performance differences between signal enhancement strategies.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 - under Grants 390895286 and 352015383 - SFB 1330 C6. We would like to thank Mathias Blau and Felix Stärz at the Jade Hochschule in Oldenburg for providing their help and expertise and the measurement setup for the high resolution HRTF set. We also would like to thank Annika Meyer-Hilberg and Janique Reinwaldt for conducting the listening experiments.

VI. References Section

References

[1] G. W. Elko and A.-T. N. Pong, "A simple adaptive first-order differential microphone," in *Proc. WASPAA*, 1995, pp. 169–172.

[2] T. de Bogaert, J. Wouters, T. J. Klasen, and M. Moonen, "Distortion of interaural time cues by directional noise reduction systems in modern digital hearing aids," in *Proc. WASPAA*, 2005, pp. 57–60.

[3] D. Marquardt and S. Doclo, "Performance Comparison of Bilateral and Binaural MVDR-based Noise Reduction Algorithms in the Presence of DOA Estimation Errors," in *Proc. ITG Symposium*, 2016, pp. 1–5.

[4] M. Tammen and S. Doclo, "Deep Multi-Frame MVDR Filtering for Binaural Noise Reduction," in *Proc. IWAENC*, 2022, pp. 1–5.

[5] H. Schröter, T. Rosenkranz, A.-N. Escalante-B, and A. Maier, "Low latency speech enhancement for hearing aids using deep filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2716–2728, 2022.

[6] T. Gajecki and W. Nogueira, "Deep latent fusion layers for binaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3127–3138, 2023.

[7] T. Goehring, X. Yang, J. J. M. Monaghan, and S. Bleeck, "Speech enhancement for hearing-impaired listeners using deep neural networks with auditory-model based features," in *Proc. EUSIPCO*, 2016, pp. 2300–2304.

[8] E. W. Healy, K. Tan, E. M. Johnson, and D. Wang, "An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 149, pp. 3943–3953, 2 2021.

[9] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Proc. Interspeech*, 2021, pp. 686–690.

[10] P. U. Diehl, Y. Singer, H. Zilly, U. Schönfeld, P. Meyer-Rachner, M. Berry, H. Sprekeler, E. Sprengel, A. Pudszuhn, and V. M. Hofmann, "Restoring speech intelligibility for hearing aid users with deep learning," *Scientific Reports*, vol. 13, pp. 1–12, 2023.

[11] P. U. Diehl, H. Zilly, F. Sattler, Y. Singer, K. Kepp, M. Berry, H. Hasemann, M. Zippel, M. Kaya, P. Meyer-Rachner, A. Pudszuhn, V. M. Hofmann, M. Vormann, and E. Sprengel, "Deep learning-based denoising streamed from mobile phones improves speech-in-noise understanding for hearing aid users," *Frontiers in Medical Engineering*, vol. 1, pp. 1–11, 2023.

[12] M. A. Akeroyd, W. Bailey, J. Barker, T. J. Cox, J. F. Culling, S. Graetzer, G. Naylor, Z. Podwińska, and Z. Tu, "The 2nd Clarity Enhancement Challenge for Hearing Aid Speech Intelligibility Enhancement: Overview and Outcomes," in *Proc. ICASSP*, 2023, pp. 1–5.

[13] S. Cornell, Z.-Q. Wang, Y. Masuyama, S. Watanabe, M. Pariente, N. Ono, and S. Squartini, "Multi-Channel Speaker Extraction with Adversarial Training: The Wavlab Submission to The Clarity ICASSP 2023 Grand Challenge," in *Proc. ICASSP*, 2023, pp. 1–2.

[14] N. L. Westhausen and B. T. Meyer, "Binaural multichannel blind speaker separation with a causal low-latency and low-complexity approach," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 238–247, 2024.

[15] ——, "Low bit rate binaural link for improved ultra low-latency low-complexity multichannel speech enhancement in Hearing Aids," in *Proc. WASPAA*, 2023, pp. 1–5.

[16] J. Karrenbauer, S. Klein, S. Schönewald, L. Gerlach, M. Blawat, J. Benndorf, and H. Blume, "SmartHeaP - A High-level Programmable, Low Power, and Mixed-Signal Hearing Aid SoC in 22nm FD-SOI," in *Proc. ESSCIRC*, 2022, pp. 265–268.

[17] K. C. Wagener, T. Brand, and B. Kollmeier, "Development and evaluation of a german sentence test part iii: Evaluation of the oldenburg sentence test," *Zeitschrift für Audiologie*, vol. 38, pp. 86–95, 1999.

[18] K. C. Wagener and T. Brand, "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters," *International Journal of Audiology*, vol. 44, pp. 144–156, 2005.

[19] G. Grimm, J. Luberadzka, and V. Hohmann, "A toolbox for rendering virtual acoustic environments in the context of audiology," *Acta Acustica united with Acustica*, vol. 105, pp. 566–578, 2019.

[20] C. Pavlovic, R. Kassayan, S. R. Prakash, H. Kayser, V. Hohmann, and A. Atamaniuk, "A high-fidelity multi-channel portable platform for development of novel algorithms for assistive listening wearables," *The Journal of the Acoustical Society of America*, vol. 146, pp. 2878–2878, 2 2019.

[21] H. Kayser, T. Herzke, P. Maanen, M. Zimmermann, G. Grimm, and V. Hohmann, "Open community platform for hearing aid algorithm research: open Master Hearing Aid (openMHA)," *SoftwareX*, vol. 17, pp. 1–10, 1 2022.

[22] J. M. Kates, "Extending the hearing-aid speech perception index (haspi): Keywords, sentences, and context," *The Journal of the Acoustical Society of America*, vol. 153, pp. 1662–1673, 2 2023.

[23] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[24] Y. Nejime and B. C. J. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *The Journal of the Acoustical Society of America*, vol. 102, pp. 603–615, 1997.

[25] Y. Luo, C. Han, and N. Mesgarani, "Ultra-Lightweight Speech Separation Via Group Communication," in *Proc. ICASSP*, 2021, pp. 16–20.

[26] H. Schröter, A. Maier, A. N. Escalante-B, and T. Rosenkranz, "Deepfilternet2: Towards Real-Time Speech Enhancement on Embedded Devices for Full-Band Audio," in *Proc. IWAENC*, 2022, pp. 1–5.

[27] Y. Luo, C. Han, and N. Mesgarani, "Group communication with context codec for lightweight source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1752–1761, 2021.

[28] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020, pp. 2492–2496.

[29] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020.

[30] T. Wendt, S. van de Par, and S. d. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *Journal of the Audio Engineering Society*, vol. 62, pp. 748–766, 11 2014.

[31] M. Blau, A. Budnik, and S. van de Par, "Assessment of perceptual attributes of classroom acoustics: real versus simulated room," in *Proc. Institute of Acoustics*, vol. 40 Pt.3, 2018, pp. 556–564.

[32] F. Stärz, L. O. H. Kroczek, S. Rosskopf, A. Mühlberger, S. van de Par, and M. Blau, "Perceptual comparison between the real and the auralized room when being presented with congruent visual stimuli via a head-mounted display," in *Proc. ICA*, 2022.

[33] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Interspeech*, 9 2019, pp. 1368–1372.

[34] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *Proc. ICASSP*, 2021, pp. 656–660.

[35] B. Jacob, K. Skirmantas, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," in *Proc. CVPR*, 2018, pp. 2704–2713.

[36] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. TSP*, 2021, pp. 72–76.

[37] P. Seetharaman, G. Wichern, B. Pardo, and J. L. Roux, "Autoclip: Adaptive Gradient Clipping for Source Separation Networks," in *Proc. MLSP*, 2020, pp. 1–6.

[38] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an international speech test signal (ists)," *International Journal of Audiology*, vol. 49, pp. 891–903, 2010, pMID: 21070124.

[39] G. Grimm and V. Hohmann, "First order ambisonics field recordings for use in virtual acoustic environments in the context of audiology," 12 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3588303

[40] D. Oetting, V. Hohmann, J.-E. Appell, B. Kollmeier, and S. D. Ewert, "Restoring perceived loudness for listeners with hearing loss," *Ear and Hearing*, vol. 39, pp. 664–678, 7 2018.

[41] N. Bisgaard, M. S. M. G. Vlaming, and M. Dahlquist, "Standard audiograms for the iec 60118-15 measurement procedure," *Trends in Amplification*, vol. 14, pp. 113–120, 2010.

[42] A. Briegleb, M. M. Halimeh, and W. Kellermann, "Exploiting spatial information with the informed complex-valued spatial autoencoder for target speaker extraction," in *Proc. ICASSP*, 2023, pp. 1–5.

[43] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 542–553, 2024.

[44] J. K. Gerald, S. Favrot, J. G. Desloge, T. M. Streeter, and C. R. Mason, "Design and preliminary testing of a visually guided hearing aid," *The Journal of the Acoustical Society of America*, vol. 133, pp. EL202–EL207, 2 2013.

[45] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Science Advances*, vol. 5, pp. 1–11, 2019.

[46] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using eeg-based auditory attention decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 862–875, 2020.