



# Subjective quality evaluation of personalized own voice reconstruction systems

Mattes Ohlenbusch<sup>1,\*</sup>, Christian Rollwage<sup>1</sup>, Simon Doclo<sup>1,2</sup>, and Jan Rannies<sup>1,2</sup>

<sup>1</sup>Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Oldenburg, Germany

<sup>2</sup>Carl von Ossietzky Universität Oldenburg, Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Oldenburg, Germany

Received 30 September 2025, Accepted 23 February 2026

**Abstract** – Own voice pickup technology for hearable devices facilitates communication in noisy environments. Own voice reconstruction (OVR) systems enhance the quality and intelligibility of the recorded noisy own voice signals. Since disturbances affecting the recorded own voice signals depend on individual factors, personalized OVR systems have the potential to outperform generic OVR systems. In this paper, we propose personalizing OVR systems through data augmentation and fine-tuning, comparing them to their generic counterparts. We investigate the influence of personalization on speech quality assessed by objective metrics and conduct a subjective listening test to evaluate quality under various conditions. In addition, we assess the prediction accuracy of the objective metrics by comparing predicted quality with subjectively measured quality. Our findings suggest that personalized OVR provides benefits over generic OVR for some talkers only. Our results also indicate that performance comparisons between systems are not always accurately predicted by objective metrics. In particular, certain disturbances lead to a consistent overestimation of quality compared to actual subjective ratings.

**Keywords.** Own voice, Hearables, Personalization, Subjective quality, Deep neural networks

## 1 Introduction

Speech communication is often impaired in noisy environments. In-the-ear hearable devices, i.e., smart earpieces with a loudspeaker and one or more microphones, can be used to improve communication in such environments, e.g., by capturing and transmitting the user's own voice to a mobile phone or another hearable [1, 2]. Here, we consider the scenario where a hearable with an outer and an in-ear microphone aims to capture the user's own voice, e.g., to be transmitted via a wireless link to another hearable or a mobile phone. The outer microphone captures environmental noise along with recording the own voice. While the in-ear microphone benefits from the attenuation of environmental noise due to ear canal occlusion, the recorded own voice suffers from low-frequency amplification (below ca. 1 kHz), band-limitation (above ca. 2 kHz), and body-produced noise [3, 4]. Own voice recorded at the in-ear microphone consists of an air-conducted and a body-conducted component. The air-conducted component strongly depends on the tightness

of the fit in the ear canal. The amount of body-conducted own voice recorded at the in-ear microphone depends on hearable device properties, such as device fit and insertion depth [5, 6], individual anatomic factors such as residual ear canal volume and shape [7, 8], the generated sounds or phonemes being uttered [9, 10], and mouth movements [11]. Environmental noise recorded at the in-ear microphone also varies with the device fit to the individual ear shape [12].

For communication applications, an own voice reconstruction (OVR) system is needed in order to reconstruct own voice from noisy hearable signals. Previous traditional signal processing OVR approaches are based on e.g., equalization filter design [13], statistical modeling [14], or non-linear bandwidth extension of in-ear own voice signals [15]. More recently proposed deep neural network (DNN)-based OVR systems are commonly designed to work for multiple potential device users [16–20]<sup>1</sup>, which is achieved by training them with data from multiple talkers in order to achieve robustness to

\*Corresponding author:

[mattes.ohlenbusch@idmt.fraunhofer.de](mailto:mattes.ohlenbusch@idmt.fraunhofer.de)

<sup>1</sup> Although some of these approaches have been proposed and validated for body-conduction microphones, they can also be applied to in-ear microphones.

individual variation. Since such systems are not specific to any particular user, in this work we refer to them as generic systems. However, since the degradations affecting noisy hearable signals are subject to several individual factors, personalized OVR systems could provide a benefit over generic systems by accounting for individual differences.

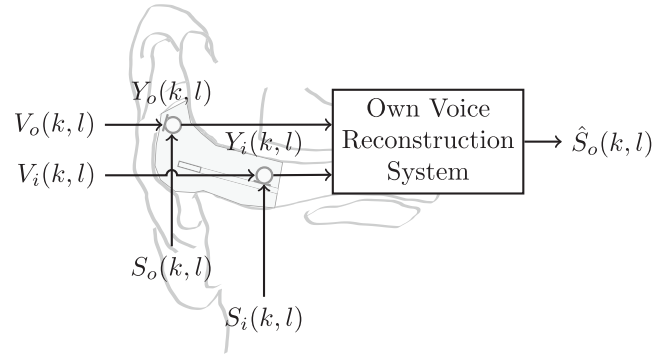
In [21], it has been proposed to personalize an OVR system by first training a generic system, and then fine-tuning the system incorporating speaker identification information into the system. By comparison, the personalized systems achieved higher reconstruction performance than generic systems in metrics predicting quality and intelligibility. Similarly, in [22] it has been proposed to personalize an OVR system for smart glasses by first pre-training a generic system for bandwidth extension of band-limited speech, and then fine-tuning the system using few recorded body-conduction signals. The personalized system achieved higher reconstruction performance in predictive metrics compared to the pre-trained generic system and compared to generic systems trained with own voice signals of multiple talkers (without pre-training). While in both [21, 22] personalization yields a benefit over generic systems, it is not yet known whether personalized training before fine-tuning can yield additional benefits in performance over pre-training a generic system. In addition, quality predictions by objective metrics often do not correlate well with subjective ratings of body-conducted speech [23]. To our knowledge, a formal subjective evaluation study of personalized OVR systems has not yet been presented.

The present study therefore extends previous work: We train generic and personalized OVR systems and compare them using objective metrics and systematic subjective ratings. We investigate both generic and personalized data augmentation for pre-training, and both generic and personalized fine-tuning. With respect to evaluation methodology, this study aims to provide insights into which instrumental metrics are most suitable for assessing OVR performance. OVR may be a particularly challenging case for performance assessment, since depending on the approach and microphone signals used, the metrics may need to be able to account for the effects of bandwidth limitation and extension.

## 2 Own voice reconstruction

### 2.1 Signal model

We consider a hearable device equipped with an outer microphone and an in-ear microphone, as depicted in Figure 1. The microphone signals are denoted by subscripts  $o$  for the outer microphone and  $i$  for the in-ear microphone. We assume that the hearable is worn by a talker in a noisy environment. In the short-time Fourier transform (STFT) domain,  $S_o(k, l)$  and  $S_i(k, l)$  denote the own voice signals of the talker at both microphones, where  $k$  and  $l$  denote the frequency index and the frame



**Figure 1.** Block diagram of own voice reconstruction using an outer and an in-ear microphone of a hearable.

index. The noisy outer and in-ear microphone signals are given by

$$Y_o(k, l) = S_o(k, l) + V_o(k, l), \quad (1)$$

$$Y_i(k, l) = S_i(k, l) + V_i(k, l), \quad (2)$$

where the noise components are denoted by  $V_o(k, l)$  and  $V_i(k, l)$ . We assume the noise components predominantly consist of environmental noise at both microphones, but also microphone self-noise with a much lower level at both microphones and additional body-produced noise at the in-ear microphone.

### 2.2 Generic and personalized own voice reconstruction systems

The goal of own voice reconstruction is to estimate the clean own voice signal  $S_o(k, l)$  from the noisy outer and in-ear microphone signals using a DNN-based OVR system  $\mathcal{D}$ , i.e.,

$$\hat{S}_o(k, l) = \mathcal{D}\{Y_o(k, l), Y_i(k, l)\}. \quad (3)$$

If the OVR system is trained to be able to reconstruct own voice of multiple talkers, we refer to the system as generic. For noisy own voice signals  $Y_o^a(k, l)$  and  $Y_i^a(k, l)$ , the same generic system  $\bar{\mathcal{D}}$  is used for any target talker  $a$ , i.e.,

$$\hat{S}_o(k, l) = \bar{\mathcal{D}}\{Y_o^a(k, l), Y_i^a(k, l)\}. \quad (4)$$

In this case, neither the system  $\bar{\mathcal{D}}$  nor its output  $\hat{S}_o$  are personalized for talker  $a$ . Training a generic OVR system requires a sufficient amount of training data from multiple target talkers. The distortions affecting the in-ear own voice signal (e.g., band-limitation, low-frequency attenuation) depend on individual factors such as fit quality and insertion depth. As a result, a generic OVR system may not achieve the same quality for all talkers.

In this work, we propose to train an OVR system to reconstruct own voice of a single target talker. For a target talker  $a$ , the corresponding personalized OVR system  $\mathcal{D}^a$  aims to reconstruct the speech of this particular talker, i.e.,

$$\hat{S}_o^a(k, l) = \mathcal{D}^a\{Y_o^a(k, l), Y_i^a(k, l)\}. \quad (5)$$

The personalized system produces a personalized output  $\hat{S}_o^a$ . However, obtaining such a personalized OVR system requires a sufficient amount of talker-specific training data.

To investigate personalized own voice reconstruction, we compare generic and personalized variants of the same DNN architecture originally proposed in [24]. The architecture is referred to as frequency- and time-domain joint nonlinear filter (FT-JNF). This architecture has previously been applied to OVR in [18, 25, 26]. In this work, we use a variant with 256 hidden units in the first and 128 hidden units in the second LSTM layer, respectively, leading to a DNN size of 466 k parameters with a complexity of 7.55 G MACs (Multiply-Accumulate operations) per second. Figure 2 shows the OVR system architecture considered in this work. Different from the original architecture in [24], we compute separate complex-valued STFT masks  $M_o, M_i$  for the outer and in-ear microphone, respectively, and the output signal is obtained as

$$\hat{S}_o(k, l) = \sum_{m \in \{o, i\}} M_m(k, l) \cdot Y_m(k, l). \quad (6)$$

Different from [24], we directly use the output of the tanh activation function as the real and imaginary parts of the masks.

### 2.3 Training-based personalization

In order to train an OVR system using both an outer and an in-ear microphone, multi-channel own voice signals are required as training data. A data augmentation method based on phoneme-specific relative transfer functions (RTFs) was proposed in [25]. The method was shown to improve OVR performance compared to only using a small recorded dataset of own voice signals directly for training.

#### 2.3.1 Generic own voice data augmentation

In [27], a method was proposed to simulate in-ear own voice signals from outer microphone signals, enabling generation of additional training utterances not included in the recorded dataset. The amount of body-conducted own voice recorded at an in-ear microphone can be observed to change depending on the generated sounds or phonemes being uttered [9, 10] and depending on mouth movements [11]. For this reason, the method proposed in [27] assumes a separate linear transfer function for each phoneme, which can be used to account for these time-varying changes. From the recorded dataset, phoneme-specific RTFs between the outer and the in-ear microphone are estimated for each recorded talker  $a$ , over all time frames in which a specific phoneme  $p$  occurs. The estimated individual, phoneme-specific RTFs are denoted by  $\hat{H}_p^a(k)$ . We refer to a set of all phoneme-specific RTFs from a single talker  $a$  as the transfer characteristics model of talker  $a$ . For RTF estimation, we assume the small

recorded dataset contains no environmental noise. Sensor noise and body-produced noise are assumed negligible compared to the own voice.

For simulation, an outer microphone signal  $S_o(k, l)$  of a random different talker is phoneme-annotated to obtain the phoneme annotation sequence  $p_o(l)$ , where  $p_o(l)$  denotes the phoneme at frame  $l$ . The simulated in-ear signal  $\hat{S}_i^a(k, l)$  of a random talker is then obtained as

$$\hat{S}_i(k, l) = \hat{H}_{p_o(l)}(k) \times S_o(k, l), \quad (7)$$

where  $\times$  denotes multiplication. Additionally, to avoid artifacts during phoneme transitions, temporal smoothing of  $\hat{H}_{p_o(l)}(k)$  is carried out (see [27] for details). Instead of assuming recorded outer microphone signals are available, it was proposed in [25] to use clean speech signals from standard datasets instead. An OVR system can then be trained with augmented own voice signals, consisting of clean speech signals used as the outer microphone own voice signal and the corresponding simulated in-ear own voice signal. Since standard datasets are readily available, this method allows for the simulation of a large amount of simulated in-ear own voice signals. In [25], a transfer characteristic model from a random talker was selected for each new utterance to obtain a simulated dataset of multiple talkers that can be used to train a generic OVR system.

#### 2.3.2 Personalized own voice data augmentation

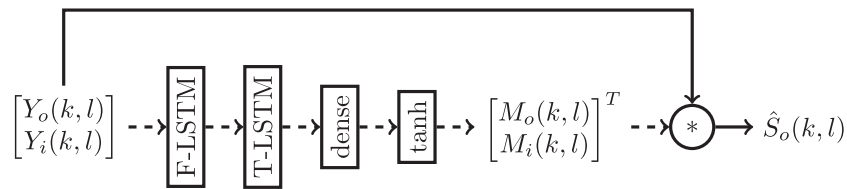
A personalized OVR system can be trained based on personalized own voice data augmentation. For this purpose, we propose to perform augmentation using only a single transfer characteristic model of a single target talker  $a$ . Similar to equation (7), the simulated in-ear signal  $\hat{S}_i^a(k, l)$  of target talker  $a$  is obtained as

$$\hat{S}_i^a(k, l) = \hat{H}_{p_o(l)}^a(k) \times S_o(k, l). \quad (8)$$

This way, an entire personalized own voice dataset for talker  $a$  can be simulated. While this data augmentation accounts for individual differences in own voice transfer characteristics, it does not account for other speech production characteristics such as prosody and pitch.

#### 2.3.3 Fine-tuning

After training an OVR system with augmented own voice signals, the recorded own voice signals can be used to fine-tune the system, further improving performance. In [25], it was shown that fine-tuning of the entire OVR system is more beneficial compared to only fine-tuning parts of the system. Fine-tuning is carried out with a smaller initial learning rate than the learning rate used for training with augmented signals. In order to fine-tune a generic system, it is possible to apply either generic or personalized fine-tuning. For *generic fine-tuning*, the recorded own voice signals of all available talkers are used. Since this procedure aims to reconstruct the own voice



**Figure 2.** DNN-based own voice reconstruction system based on the FT-JNF architecture from [24]. The first LSTM layer operates in frequency direction (F-LSTM) and the second LSTM layer operates in time direction (T-LSTM). Figure adapted from [18].

of multiple individual talkers, the fine-tuned OVR system is generic. For *personalized fine-tuning*, the recorded own voice signals of only the target talker are used. Since this procedure aims to reconstruct the own voice of a single individual talker, the fine-tuned OVR system is personalized.

### 3 OVR system training setup

An almost identical training setup as described in [25] was used. The split of the dataset of recorded own voice signals is different (see Sect. 3.1), and in addition to generic data augmentation and fine-tuning, personalized data augmentation and fine-tuning are also considered.

#### 3.1 Datasets

The evaluation uses clean recorded own voice signals made with the Hearpiece hearable prototype [28]. A dataset of German own voice signals of 18 talkers with 306 utterances each is split into 206 utterances for training, 50 utterances for validation, and 50 utterances for testing, respectively. For personalized training, all training utterances of the target talker are used in data augmentation or fine-tuning. For generic training, a random subset consisting of 206 utterances is selected from the training utterances of all 18 talkers, so that the number of recorded own voice signals is the same for generic and personalized training. The augmented own voice signals are obtained by augmenting 10% of the German portion of the CommonVoice dataset [29] (v11.0), which corresponds to 115.7 hours of speech signals.

The noise signals at both microphones used for training and evaluating the OVR systems based on instrumental metrics are a spatialized version of the fifth DNS challenge [30], obtained following the procedure in [18] using individually matched, measured transfer functions for the same users as in the dataset of recorded own voice signals. Measurements from 8 horizontal directions in 45°-steps at a distance of 1.5 m are used to compute either point source signals (single direction) or pseudo-diffuse noise signals (8 directions). For obtaining the noisy input signals, the spatialized outer and in-ear microphone noise signals are added to the augmented or recorded own voice signals at both microphones.

#### 3.2 Evaluation details

The evaluation was carried out similarly to [25]. For evaluation, OVR performance was evaluated at fixed signal-to-noise ratios (SNRs) of  $-10$ ,  $-5$ ,  $0$ ,  $5$ , and  $10$  dB, and results were averaged over test set examples and the considered SNRs. The performance was assessed in terms of instrumental metrics (see Sect. 4). In order to investigate the influence of personalized OVR, the performance of generic and personalized systems was compared for each talker in the recorded dataset. For generic systems, a single system was trained with noisy own voice signals from all 18 talkers. The generic system was then evaluated separately on each talker’s test set (excluding training and validation utterances). For personalized systems, a separate personalized system was trained for each talker. Each personalized system was then evaluated only on the test set of the same corresponding talker, respectively.

### 4 Objective quality prediction metrics

Since it is currently unknown which metrics are best suited for predicting OVR performance, a variety of instrumental metrics are tested against subjective ratings. OVR systems are commonly evaluated using speech enhancement metrics, which aim to predict speech quality, intelligibility, or listening effort. Some of the metrics require a reference signal (intrusive metrics), which is a clean speech signal corresponding to the speech content in the noisy or processed signal to be evaluated by the metric. In this work, the following intrusive metrics are investigated:

**PESQ** Wideband perceptual evaluation of speech quality (PESQ) [31] is a metric predicting speech quality. Although it is often used beyond its original scope and has since been superseded by POLQA [32], it remains a popular metric to evaluate speech enhancement systems. In particular, in [33] it was observed that PESQ predictions correlate well with subjective ratings of speech recorded at an in-ear microphone, and slightly better than POLQA. When predicting quality of bandwidth-extended signals in [34] PESQ showed low correlations, although slightly better than POLQA. The scale of PESQ output values ranges from 0.5 to 4.5.

**ESTOI** Extended short-time objective intelligibility (ESTOI) was originally designed to predict intelligibility. Strictly, this requires a score transformation based on the evaluation material [35]. Nevertheless, it is often used without transformation to evaluate the performance of speech enhancement systems. In Haurer et al. [17], the original STOI [36] was observed to predict subjective quality ratings well. The extended STOI is designed to work for a wider scope than the original STOI, including highly modulated noise. The scale of raw ESTOI output values ranges from 0 to 1.

**GPSM<sup>q</sup>** The generalized power spectrum model for quality GPSM<sup>q</sup> [37] uses signal-to-noise ratios in the power and envelope power domains to model the addition or removal of energy by processing. This metric was designed to predict the quality introduced by linear and non-linear distortions, making it suitable for assessing e.g., audio codecs, noise reduction, or source separation systems. In this study, we use the raw objective perceptual measure (OPM) without perceptual thresholding and resample input signals to 32 kHz. The scale of predicted values ranges from 0 to 17.

**eMoBi-Q** The efficient model for binaural audio quality (eMoBi-Q) [38] is a simplified version of MoBi-Q [39], providing a lean structure and a new, compact binaural path. The monaural path captures spectral distortions, but not nonlinear distortions. While eMoBi-Q simplifies some computational aspects, accuracy for predicting quality in hearing device applications is maintained. The scale of predicted values by eMoBi-Q ranges from 0 to 1.

**PEMO-Q PSM** An audio quality model which is based on a psychoacoustically validated auditory perception model (PEMO) [40] is PEMO-Q [41]. In this study, its most basic metric, i.e., the Perceptual Similarity Measure (PSM), is used. PSM is the linear cross correlation between the internal representations of a pair of test and reference signals. The internal representations are computed by the perception model. In this case, the simpler variant is used in which internal representations are obtained by modeling amplitude modulation processing with a low-pass filter. A voice activity detector is used to select signal segments containing speech before processing the signal by the model. Being a Pearson correlation coefficient, PSM can assume values between  $-1$  and  $1$ , with  $1$  indicating perceptual identity between test and reference signal (interpreted as maximum quality of the test signal); practically, output values are in the range between 0 and 1.

**SCOREQ distance** Speech contrastive regression for quality (SCOREQ) [42] is a method for predicting speech quality designed to capture the continuous nature of the MOS scale. When applied to signals for which a reference signal is available, a distance value can be computed indicating similarity to the reference signal. The distance ranges from 0 to infinity, with smaller values indicating higher similarity.

In this work, the reference signal for all intrusive metrics is the clean outer microphone signal. A higher value indicates better quality for all metrics except SCOREQ distance, where a lower distance value indicates better quality. In contrast to intrusive metrics, non-intrusive metrics do not require a reference signal. Many popular non-intrusive metrics are based on deep learning and are trained to predict subjective ratings from noisy or processed signals only. In this work, the following non-intrusive metrics are investigated:

**DNSMOS** Deep Noise Suppression Mean Opinion Score (DNSMOS) [43] is a non-intrusive DNN-based metric to predict quality of an audio signal in terms of speech/signal quality (DNSMOS SIG), background quality (DNSMOS BAK), and overall quality (DNSMOS OVRL) based on P.835 [44], as well as overall quality based on P.808 [45], here referred to as DNSMOS P808. In both cases, DNSMOS predicts values on the MOS scale from 1 to 5.

**SCOREQ MOS** different from the SCOREQ distance, the SCOREQ model can also be used without a reference signal to predict values on the MOS scale from 1 to 5 [42].

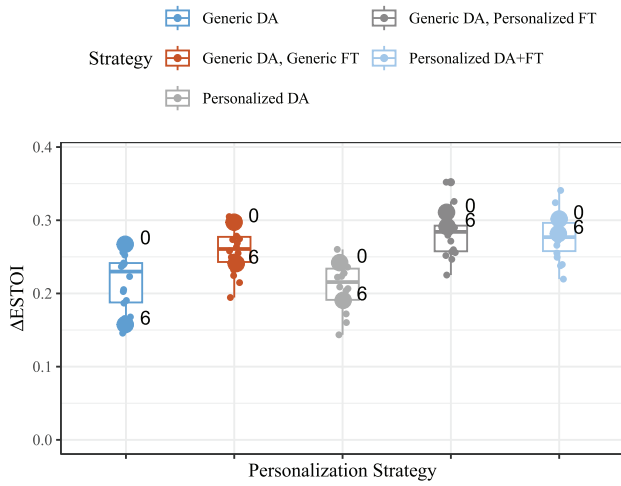
**WV-MOS** Band-limited and bandwidth-extended signals are hard to evaluate in terms of quality using objective metrics designed for full-band signals [23, 34]. To address this, it was proposed in [46] to predict MOS ratings using a DNN-based approach based on a pretrained wav2vec2.0 (WV-MOS). WV-MOS predicts values on the MOS scale from 1 to 5.

**LEAP** The model for Listening Effort prediction from Acoustic Parameters (LEAP) [47, 48] is based on an automatic phoneme recognizer for German speech. Speech degradations, such as additive noise, distortions, or reverberation, increase the uncertainty of the recognition process, which has been found to correlate very well with human ratings of perceived listening effort. LEAP predicts the perceived listening effort on a subjective scale, ranging from 1 (“no effort”) to 13 (“extreme effort”) and 14 (“only noise” or “no speech perceivable”) [49].

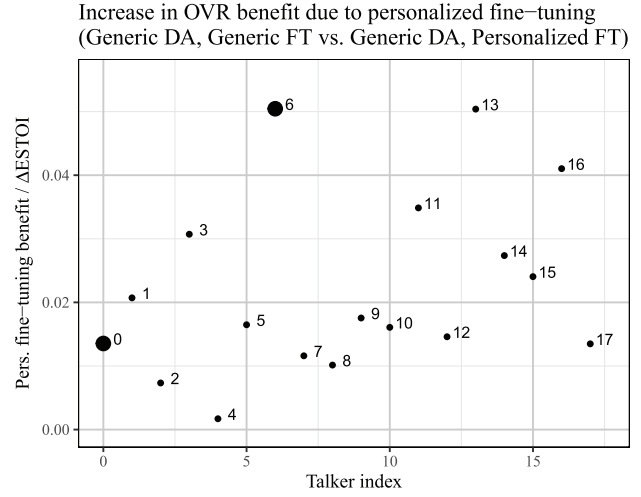
## 5 Results of instrumental assessment

Figures 3a and 4a show the improvements ( $\Delta$ ) in ESTOI and PESQ, respectively, compared to the noisy outer microphone signals.

The results for individual target talker test sets are shown both as individual points as well as boxplots over all target talkers per personalization strategy. Two exemplary target talkers (Numbers 0 and 6) are highlighted. All strategies consistently improved both metrics across all talkers compared to the noisy outer microphone signals. Strategies that include a fine-tuning step perform better than those without. Personalized data augmentation without fine-tuning leads to slightly lower scores

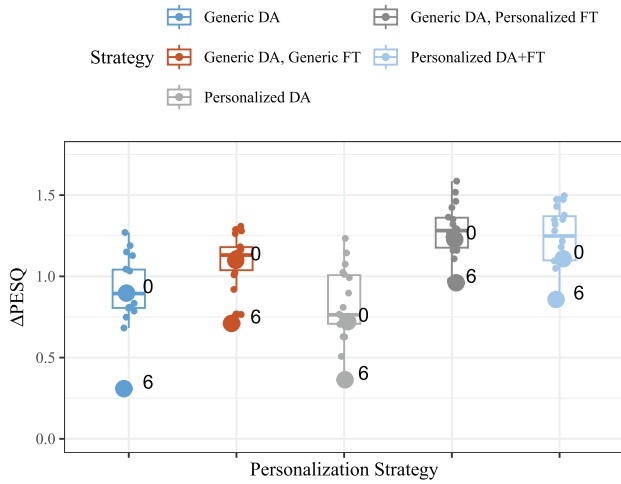


(a) ESTOI improvement.

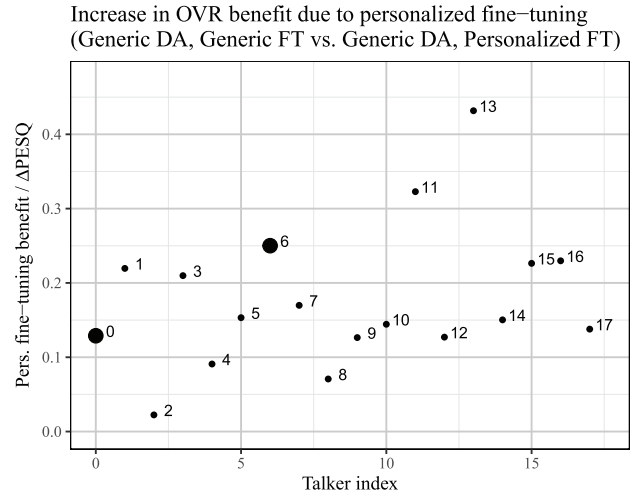


(b) Predicted benefit from personalized fine-tuning.

**Figure 3.** (a) ESTOI improvement achieved by OVR systems trained with different personalization strategies in data augmentation (DA) and fine-tuning (FT) and (b) individual difference in ESTOI improvement between the conditions generic DA, generic FT and generic DA, personalized FT. Individual data points denote the average over individual target talker test sets. The data points labeled with 0 and 6 correspond to the talkers selected for the listening experiment as the *low predicted benefit* and *high predicted benefit*, respectively.



(a) PESQ improvement.



(b) Predicted benefit from personalized fine-tuning.

**Figure 4.** (a) PESQ improvement achieved by OVR systems trained with different personalization strategies in data augmentation (DA) and fine-tuning (FT) and (b) individual difference in PESQ improvement between the conditions generic DA, generic FT and generic DA, personalized FT. Individual data points denote the average over individual target talker test sets. The data points labeled with 0 and 6 correspond to the talkers selected for the listening experiment as the *low predicted benefit* and *high predicted benefit*, respectively.

than generic fine-tuning. When generic data augmentation is used, personalized fine-tuning leads to better performance than generic fine-tuning. Systems trained with generic data augmentation and then personalized fine-tuning outperformed those trained with personalized data augmentation and fine-tuning.

The results predict a consistent benefit of OVR systems personalized by fine-tuning over generic systems. In contrast, no consistent benefit is predicted for personalized data augmentation, which could be due to

the decrease in variance in the training data relative to generic data augmentation (18 times the number of RTFs used for augmentation). In order to investigate whether the instrumental predictions are accurate, a subjective listening experiment is performed in the following.

The benefit achieved by personalized over generic processing is different between individual target talkers. In Figures 3 and 4, two example target talkers are highlighted. For assessing the benefit of personalized over generic fine-tuning, the difference between the  $\Delta$ ESTOI

and  $\Delta$ PESQ scores in [Figures 3a](#) and [4a](#) for the conditions Generic DA, Generic FT and Generic DA, personalized FT are shown in [Figures 3b](#) and [4b](#). In case of talker 0, only a small additional improvement in ESTOI and PESQ is achieved by personalized fine-tuning when compared to the generic cases. In contrast, in case of talker 6, a large additional improvement is achieved by personalized fine-tuning in terms of ESTOI. While PESQ predicts a different ranking order for this talker than ESTOI, it still indicates an improvement of 0.25  $\Delta$ PESQ from personalized fine-tuning compared to generic fine-tuning. To reduce the measurement time for the listening experiment, we chose to investigate only the performance for these two target talkers when comparing predicted quality to subjective quality ratings. We refer to processed signals of talker 0 as the *low predicted benefit* case, and processed signals of talker 6 as the *high predicted benefit* case.

## 6 Listening experiment

### 6.1 Evaluation procedure

In order to investigate subjective quality achieved by OVR processing, a listening experiment based on the multiple stimuli with hidden reference and anchor (MUSHRA) standard for assessing intermediate audio quality [50] was conducted. Unlike the MUSHRA standard, the MUSHRA-like experiment used in this paper employed different anchor signals (see [Sect. 6.3](#)) than the lowpass-filtered clean speech signals defined in International Telecommunications Union (ITU) [50]. The experiment was carried out using the WebMUSHRA framework [51]. Participants were instructed to rate the overall quality of each signal presented to them. Before conducting the actual experiment, a training screen was presented to the participants for familiarization. The experiment was conducted in sound-proof listening booths, and stimuli (noisy and processed own voice) were presented over open-back headphones (Sennheiser HDA 650, calibrated for 70 dB SPL output).

### 6.2 Participants

Twenty-five normal-hearing native German-speaking participants (13 female, 12 male), aged  $23.8 \pm 3.8$  years (mean  $\pm$  standard deviation), took part in the listening experiment. All participants had pure-tone thresholds  $\leq 20$  dB hearing level at audiometric frequencies from 125 Hz to 8 kHz on both ears. One participant was excluded from the evaluation for not correctly identifying the reference signal in all the MUSHRA screens. All participants received hourly compensation and gave informed consent for their participation in the experiments. The methods of the experiment were approved by the ethics committee of the Carl von Ossietzky University of Oldenburg (protocol Drs.EK/2019/073-2).

### 6.3 Processing conditions

The subjective evaluation was conducted by evaluating quality of noisy and processed own voice signals in different processing conditions, noise types, and talkers. The considered processing conditions were:

**Noisy OM** The noisy, unprocessed outer microphone signal (anchor). Since the results in [Section 5](#) suggest a consistent improvement by OVR systems over the noisy unprocessed outer microphone signals, they were used as anchor signals for the MUSHRA-like listening test (instead of, e.g., clean band-limited signals).

**Noisy IM** The noisy, unprocessed in-ear microphone signal.

**MWF** An implementation of the multi-channel Wiener filter (MWF) [52] with multi-channel speech presence probability and recursive smoothing-based power spectral density estimation [53, 54] using the noisy outer and in-ear signals as input signals and using the outer microphone as the reference microphone. MWF is used as a baseline method in the subjective evaluation.

**EBEN** The extreme bandwidth extension network (EBEN) [17] is a time-domain DNN-based system. For the listening test, it was retrained with generic data augmentation and generic fine-tuning to estimate clean outer microphone own voice signals from noisy in-ear microphone signals. Training was performed using the same hyperparameters as the FT-JNF systems, including the loss function (different from the generative adversarial network training paradigm in [17]). EBEN is used as a baseline method in the subjective evaluation.

**Generic DA** The FT-JNF architecture, trained using generic data augmentation only.

**Generic DA, generic FT** The FT-JNF architecture, trained using generic data augmentation and generic fine-tuning.

**Generic DA, personalized FT** The FT-JNF architecture, trained using generic data augmentation and personalized fine-tuning.

**Personalized DA, personalized FT** The FT-JNF architecture, trained using personalized data augmentation and personalized fine-tuning.

**Reference** The clean own voice signal at the outer microphone (hidden reference).

Processing conditions were applied to own voice signals of two talkers selected based on the results presented in [Section 5](#). One talker represented a low predicted personalization benefit, the other a high predicted benefit. For each talker, own voice signals were mixed with recorded noise signals (different from the spatialized noise signals mentioned in [Sect. 3.1](#), which were used for the evaluation with instrumental metrics). Four different noise types were considered: surgery noise, metal grinder noise, pseudo-diffuse surgery noise, and pseudo-diffuse factory noise. For the surgery noise and metal grinder noise, a

noise recording from a real acoustic environment was played back from a loudspeaker 1.5 m in front of each talker while they were wearing the Hearpiece devices. For the pseudo-diffuse surgery noise and the pseudo-diffuse factory noise, time-shifted recordings from real acoustic environments were played back from 8 loudspeakers in 45°-steps with 1.5 m distance to create an approximately diffuse acoustic environment. This recording scenario matches the spatial setup of the spatialized noise signals in Section 3.1. Different from the spatialized setup, these noise signals are recorded and reflect realistic scenarios in which hearables could be used for own voice pickup [1, 2, 55]. Spectrograms of the recorded noise signals at the outer microphone are shown in Figure 5. The recorded noise signals were also used for evaluation in [18] and were only used for testing.

The resulting noisy own voice signals were mixed to an SNR of 0 dB at the outer microphone. This ensured a presentation level of the noisy own voice signal corresponding to 70 dB sound pressure level (SPL). For each combination of processing condition, talker, and noise type, utterances of three different sentences were presented to the participants, and the resulting subjective ratings were averaged over utterances.

## 7 Results of listening experiment

### 7.1 Subjective quality ratings

Figure 6 shows the subjective quality ratings for noisy and processed speech in the *low predicted benefit* case. Different subplots show the results for speech in different noise types. The ratings of each condition were very similar across noise types. In all cases, the clean reference was identified correctly. The noisy outer microphone signals were rated very low, as expected for the anchor signal, while the noisy in-ear microphone signals were rated close to 50 out of 100 on average. MWF and EBEN both performed worse than the noisy in-ear microphone signals. EBEN in particular was rated similar to the noisy outer microphone signals, indicating no quality improvement. All OVR systems trained with the proposed personalization strategies led to considerable improvements over the noisy microphone signals. In particular, the approaches using personalized fine-tuning achieved the highest quality ratings, closely followed by generic DA, generic FT.

Figure 7 shows the subjective quality ratings for noisy and processed speech in the *high predicted benefit* case. The results are very similar to those in the *low predicted benefit* case. Again, only small differences were observed between systems using generic data augmentation and generic fine-tuning and systems using personalized fine-tuning. Different from those results, the noisy in-ear microphone was rated slightly better than for speech of the target talker with low predicted benefit. In addition, speech processed by the OVR system

trained with generic data augmentation without fine-tuning were rated similar to the noisy in-ear microphone signals.

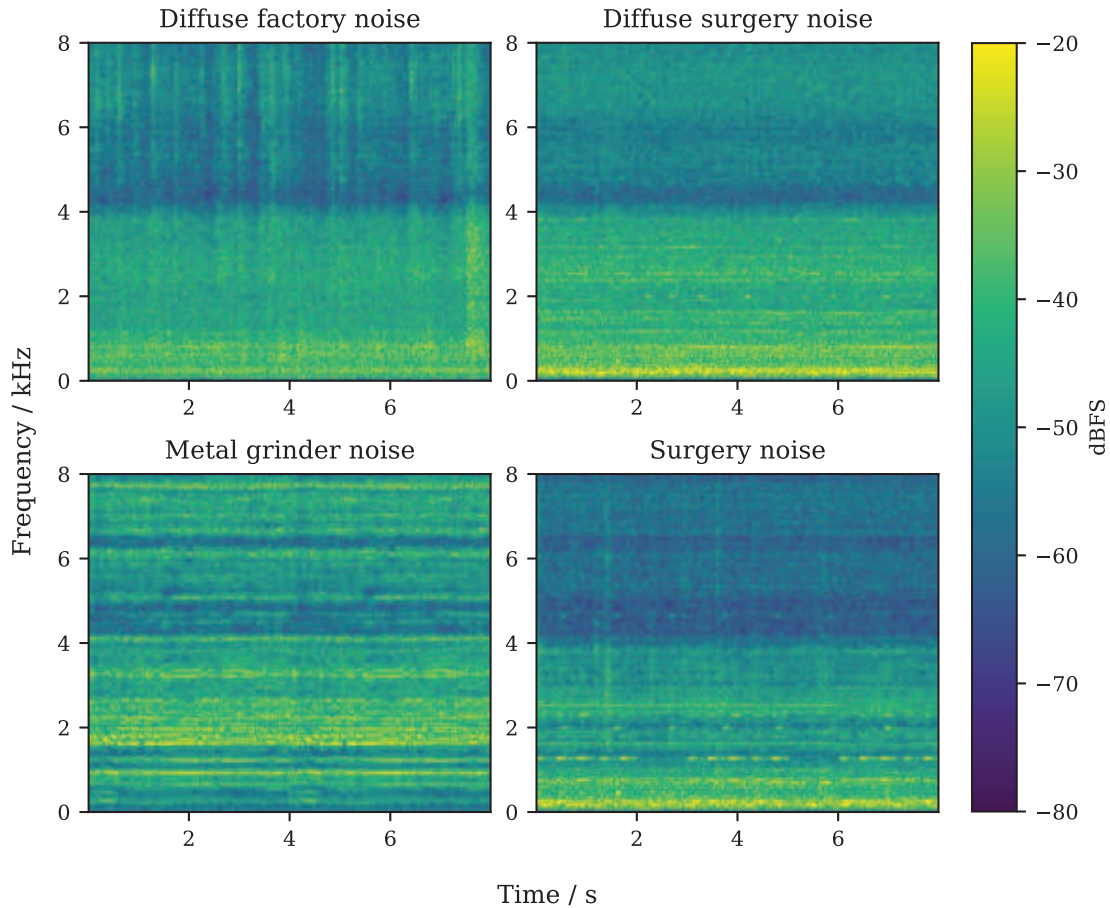
In summary, there is a consistent improvement of speech quality by the proposed personalization strategies that include fine-tuning, compared to both noisy microphone signals and baseline systems. The difference observed between the *low predicted benefit* and *high predicted benefit* cases in Section 5 was not observed in the subjective ratings. The OVR systems personalized either by data augmentation or fine-tuning do not yield substantially higher subjective ratings than the generic systems.

### 7.2 Statistical analysis of subjective ratings

Statistical inference was conducted using the R environment [56]. For statistical analysis, subjective ratings were averaged over sentences and noise types. Shapiro–Wilk tests revealed non-normality of the data. Visual inspection of QQ-plots indicated saturation effects at both ends of the MUSHRA scale. For these reasons, the analysis was carried out by means of non-parametric one-way repeated-measures analyses using separate Friedman rank sum tests for each talker considered. Wilcoxon signed rank tests were conducted as post-hoc tests, and Bonferroni-corrected for multiple comparisons. The hidden reference ratings were excluded from the statistical analysis. They served only to anchor participants and to ensure instructions were followed, and thus were not considered a processing condition.

#### 7.2.1 Low predicted benefit

In the *low predicted benefit* case, the Friedman test revealed significant differences between processing conditions ( $\chi^2(7) = 161.2, p < 0.001$ ). Post-hoc pairwise comparisons with Bonferroni correction were carried out. The resulting  $p$ -values are reported in Table 1. The post-hoc tests revealed that the noisy outer microphone signal was rated significantly worse than all other conditions except EBEN, which received similar ratings as the noisy outer microphone signal ( $p > 0.999$ ). The MWF was rated significantly higher than EBEN ( $p < 0.001$ ), but not significantly higher from the noisy in-ear microphone ( $p > 0.999$ ). All generic and personalized approaches based on FT-JNF were rated significantly higher than both noisy microphone signals and significantly higher than both of the baseline systems (MWF and EBEN). The system trained with generic data augmentation without fine-tuning was rated significantly worse than the conditions that include fine-tuning. Between the conditions that include fine-tuning, the system trained with generic data augmentation and personalized fine-tuning did not perform significantly better than the system trained with generic data augmentation and generic fine-tuning ( $p > 0.999$ ).



**Figure 5.** Spectrograms of the recorded noise signals used in the subjective evaluation.

### 7.2.2 High predicted benefit

In the *high predicted benefit* case, the Friedman test also revealed significant differences between processing conditions ( $\chi^2(7) = 151.34, p < 0.001$ ). Posthoc pairwise comparisons with Bonferroni correction were carried out. The resulting  $p$ -values are reported in Table 2. Similar to the results in the *low predicted benefit* case, here the noisy outer microphone signal was rated significantly worse than all other processing conditions except EBEN ( $p > 0.999$ ). The MWF was rated significantly higher than EBEN ( $p < 0.001$ ), but not significantly higher from the noisy in-ear microphone signal ( $p > 0.999$ ). All generic and personalized approaches based on FT-JNF were rated significantly higher than both noisy microphone signals and significantly higher than both of the baseline systems (MWF and EBEN), except for the system trained with generic data augmentation and without fine-tuning, which was not rated significantly different from the noisy in-ear microphone signals ( $p > 0.999$ ). The systems trained with fine-tuning were rated significantly higher than the system trained with generic data augmentation and without fine-tuning. Between the fine-tuned systems, there were no significant differences.

### 7.3 Prediction of subjective quality ratings

This section investigates how well the objective metrics described in Section 4 predict the subjective ratings described in Section 7.1. The prediction performance was measured in terms of correlation and root mean squared error (RMSE) between objective metric prediction and the median subjective quality rating. In particular, the Pearson linear correlation coefficient  $r$  was used to assess the accuracy of the predictions, and the Spearman rank coefficient  $\rho_S$  was used to assess the monotonicity of the predictions. The same audio signals as in the subjective evaluation were used to compute predictions for each individual signal. For each of the eight processing conditions, 24 predictions (two cases, three sentences, four noise types) were compared with the corresponding subjective ratings (median over subjects). We computed RMSE values for raw predictions (see e.g., [57]). In addition, RMSE was computed after fitting a third-order polynomial (RMSE<sub>3</sub>) to account for systematic variation in subjective ratings. For computing RMSE values, the predictions were scaled to the MUSHRA range to facilitate comparisons between metrics<sup>2</sup>. For PEMO-Q, an output

<sup>2</sup> It should be noted that some of the DNN-based metrics produced values outside their respective range, e.g., WV-MOS predictions are supposed to lie on the MOS scale between 1 and 5, but negative predictions were observed.

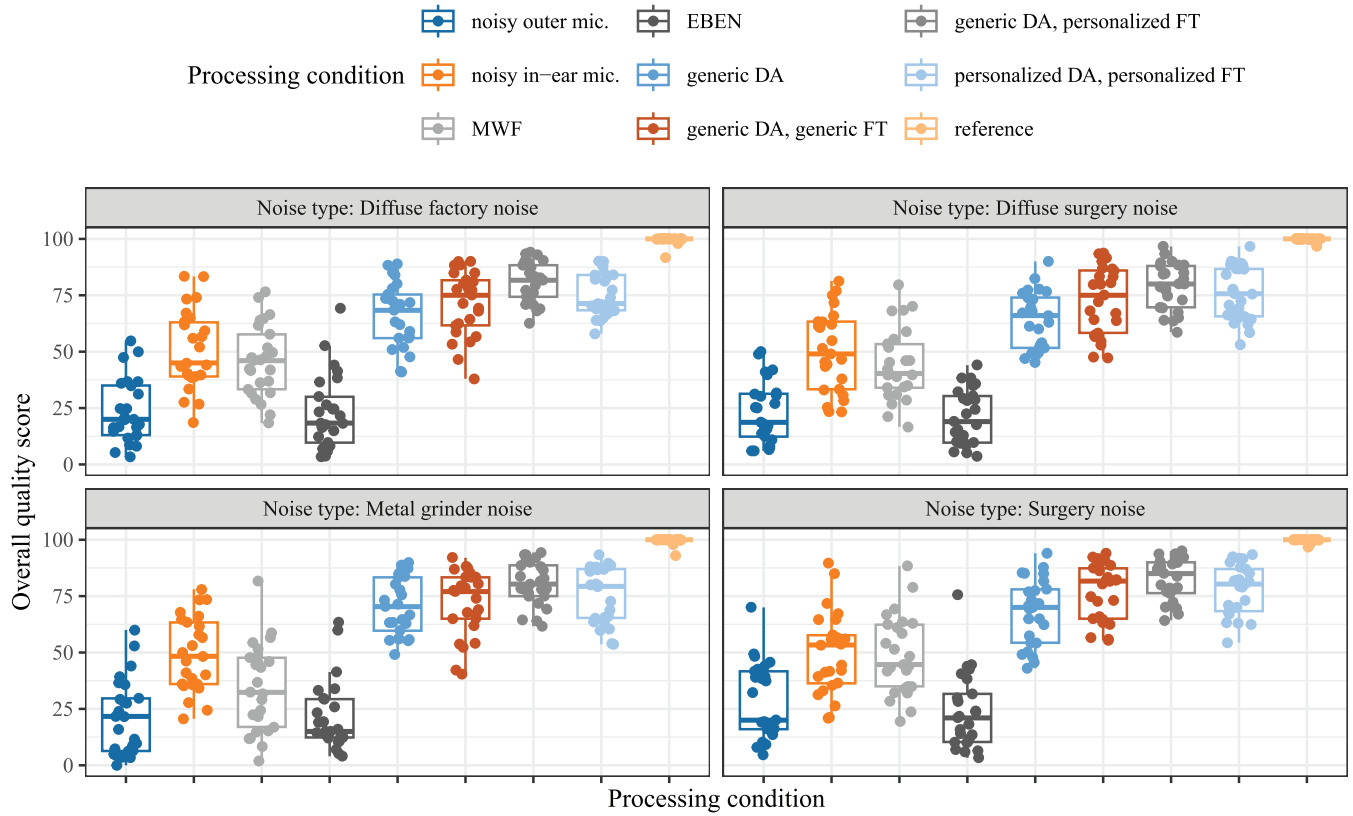


Figure 6. Subjective MUSHRA quality ratings (averaged over sentences) for speech in the *low predicted benefit* case.

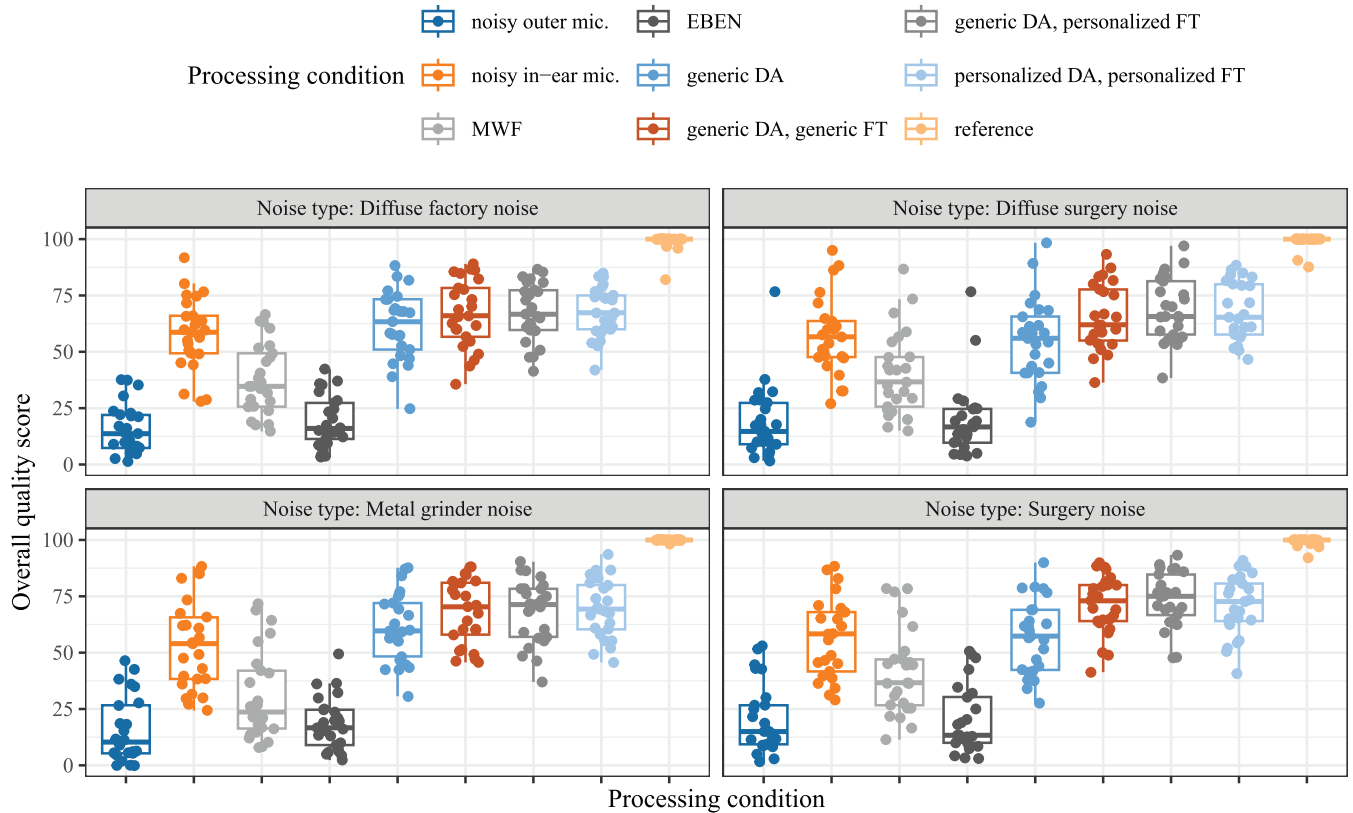


Figure 7. Subjective MUSHRA quality ratings (averaged over sentences) for speech in the *high predicted benefit* case.

**Table 1.** Resulting  $p$ -values from Wilcoxon signed rank test (Bonferroni-corrected) for the *low predicted benefit* case. Asterisks indicate significant differences.

	Noisy outer	Noisy in-ear	MWF	EBEN	gen. DA	gen. DA, gen. FT	gen. DA, pers. FT
Noisy in-ear	<0.001*	–	–	–	–	–	–
MWF	<0.001*	>0.999	–	–	–	–	–
EBEN	>0.999	<0.001*	<0.001*	–	–	–	–
gen. DA	<0.001*	0.002*	<0.001*	<0.001*	–	–	–
gen. DA, gen. FT	<0.001*	<0.001*	<0.001*	<0.001*	0.001*	–	–
gen. DA, pers. FT	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	–
pers. DA, pers. FT	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	>0.999	<0.001*

**Table 2.** Resulting  $p$ -values from Wilcoxon signed rank test (Bonferroni-corrected) for the *high predicted benefit* case. Asterisks indicate significant differences.

	Noisy outer	Noisy in-ear	MWF	EBEN	gen. DA	gen. DA, gen. FT	gen. DA, pers. FT
Noisy in-ear	<0.001*	–	–	–	–	–	–
MWF	<0.001*	<0.001*	–	–	–	–	–
EBEN	>0.999	<0.001*	<0.001*	–	–	–	–
gen. DA	<0.001*	>0.999	<0.001*	<0.001*	–	–	–
gen. DA, gen. FT	<0.001*	0.001*	<0.001*	<0.001*	<0.001*	–	–
gen. DA, pers. FT	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	>0.999	–
pers. DA, pers. FT	<0.001*	<0.001*	<0.001*	<0.001*	0.005*	>0.999	>0.999

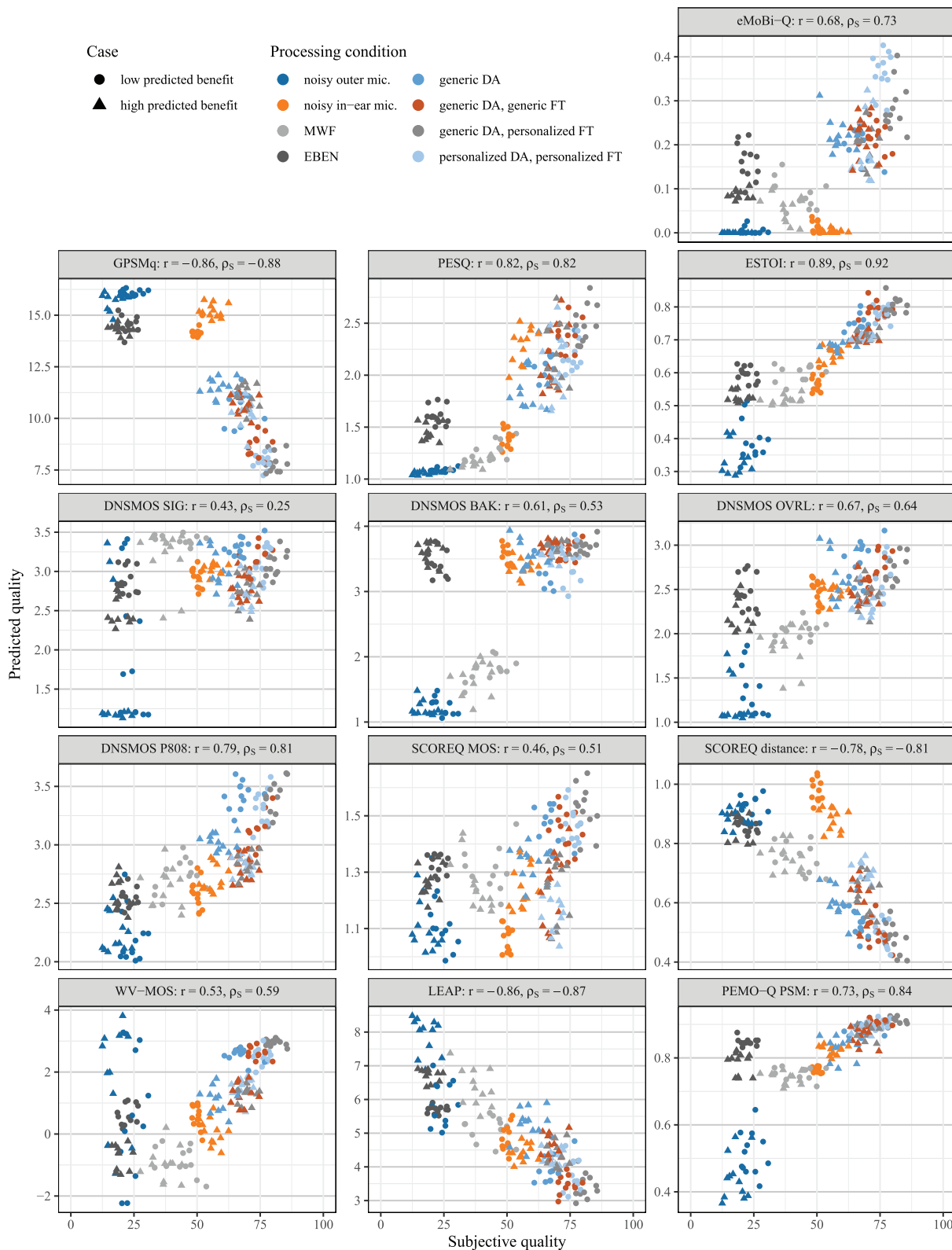
range between 0 and 1 was considered. For LEAP, an output range between 1 and 13 was considered. For SCOREQ distance, no RMSE value is reported, since the distance scale is open-ended. This scaling procedure assumes a linear relationship between the (relative) MUSHRA scale and the (absolute) scales of the metrics, such as the MOS scale.

The prediction performance of the objective metrics is shown in Figure 8. Different symbols correspond to the own voice signals of different talkers, while different colors correspond to different processing conditions. For each metric, the quality predictions of different processing conditions can be observed to group in clusters. Most metrics predicted higher quality for conditions with higher subjective ratings. In the case of DNSMOS SIG and DNSMOS BAK, there is a high spread of ratings that does not seem to follow a strong common trend. In the case of SCOREQ distance, there appears to be a strong, but negative correlation, since for the distance-based metric lower values indicate higher similarity. The SCOREQ distance values for the noisy in-ear microphone signals are similar to the noisy outer microphone signals or the signals processed by EBEN, despite the noisy in-ear microphone signals receiving much higher subjective ratings.

LEAP achieved a strong negative correlation, indicating that lower listening effort corresponds to higher subjective quality. The majority of the metrics consistently overestimated the quality of the signals processed by EBEN with respect to the other processing conditions (see clusters of black symbols in each panel). In contrast, the quality of the noisy in-ear signals was consistently underestimated by eMoBi-Q, GPSM<sup>a</sup>, and

the SCOREQ MOS and distance predictions. The bandwidth extension-based metric WV-MOS strongly overestimated the quality of the full-bandwidth noisy outer microphone signals. Comparing the predictions for low and high predicted benefit, the predictions by all metrics tend to cluster together for both cases, with the individual processing conditions forming distinct clusters. This indicates that the processing condition has a larger influence on the predictions than the case (low or high predicted benefit), which was consistently observed in the experiments as well as in most predictions.

Correlation coefficients as well as root mean squared error (RMSE) values are reported in Table 3. In terms of linear correlation, ESTOI predictions achieved the highest (absolute) correlation coefficient, followed by LEAP, GPSM<sup>a</sup>, PESQ, PEMO-Q PSM, DNSMOS P808, and SCOREQ distance. The absolute correlation achieved by SCOREQ distance is almost as high as for P808. Slightly lower correlations were achieved by eMoBi-Q and DNSMOS OVRL. DNSMOS BAK, WV-MOS, SCOREQ MOS and DNSMOS SIG achieved the lowest correlations out of the metrics considered. In terms of rank correlation, the order of metrics is similar, with the correlation of ESTOI predictions being particularly high ( $\rho_S = 0.92$ ). For most metrics, the linear and rank correlations are very close, except for DNSMOS SIG where a much lower rank correlation ( $\rho_S = 0.25$ ) was achieved than linear correlation ( $r = 0.43$ ). In terms of RMSE, the residual errors of the predicted values are generally higher compared to the RMSE values reported in, e.g., [38]. As visible in Figure 8, many of the metrics do not predict close to the upper end of their respective



**Figure 8.** Scatterplots comparing median subjective quality ratings (abscissae) and predicted objective quality (ordinates) for noisy and reconstructed own voice signals.

scale for processing conditions that were rated to have very high subjective quality or do not reach the lower end of their respective scale for processing conditions rated

to have very low subjective quality. The lowest RMSE was achieved by ESTOI, while the highest RMSE was achieved by WV-MOS. The RMSE<sub>3</sub> values after fitting

**Table 3.** Correlation coefficients and RMSE values for the considered metrics.

Metric	$r$	$\rho_S$	RMSE	RMSE <sub>3</sub>
eMoBi-Q	0.68	0.73	40.8	15.2
GPSM <sup>a</sup>	-0.86	-0.88	39.9	10.9
PESQ	0.82	0.82	23.5	11.8
ESTOI	0.89	0.92	16.0	8.7
DNSMOS SIG	0.43	0.25	20.7	18.4
DNSMOS BAK	0.61	0.53	19.8	16.8
DNSMOS OVRL	0.67	0.64	25.2	15.6
DNSMOS P808	0.79	0.81	16.7	12.5
SCOREQ MOS	0.46	0.51	49.2	18.6
SCOREQ dist.	-0.78	-0.81	-	11.9
WV-MOS	0.53	0.59	58.6	16.7
LEAP	-0.86	-0.87	36.2	10.3
PEMO-Q PSM	0.73	0.84	31.3	12.6

a third-order polynomial are generally lower than the RMSE values computed from the raw predictions. In particular, the values for ESTOI are the lowest, followed by LEAP, GPSM<sup>a</sup>, PESQ, and SCOREQ distance.

## 8 Discussion

### 8.1 Comparison to previous research

Previous research on generic OVR systems has compared different approaches using subjective ratings, e.g., [17, 19]. In [19], objective metrics (PESQ and STOI) and subjective ratings showed an improvement through OVR processing. While in [19] objective metrics and subjective ratings were not compared in terms of correlation, a comparison of objective metrics and subjective ratings in Hauret et al. [17] revealed low correlation of PESQ and SI-SDR (a technical metric) with subjective ratings, but higher correlations were observed for STOI and Noresq-MOS (a reference-free metric). However, the experiments in [17] were carried out using only simulated own voice signals and only generic systems were evaluated. In comparison, the results presented in this paper show a higher correlation for PESQ with subjective quality ratings than in [17], while ESTOI achieved similarly high correlations as STOI in [17].

In both [17, 19], OVR was only performed using single-channel body-conduction sensor signals as input, whereas the multi-channel OVR systems in this work use both an in-ear and an outer microphone as input. Additionally, personalized OVR systems are considered, and recorded own voice signals are used for evaluation.

Although previous research has already investigated personalization of OVR systems, there has not been any subjective evaluation of personalized OVR systems as far as we know. In [21], personalized OVR systems using an in-ear microphone were proposed, but only evaluated in terms of objective metrics. Due to the proposed personalization method, it was observed that the generalization to

talkers not in the training data was poor. Similarly, in [22] personalized OVR systems were proposed, but the benefit of personalization was not assessed in terms of subjective ratings. In [58], personalization was achieved by calibrating an OVR system with a few minutes of recorded speech signals from the target talker, but the benefit was not assessed in terms of subjective ratings either. This paper addresses these knowledge gaps by comparing personalized OVR systems using both subjective and objective evaluations.

In terms of personalized data augmentation, previous work has already investigated data augmentation based on text-to-speech systems for synthesizing training data for training personalized speech enhancement systems, e.g., in [59, 60]. These approaches also perform personalized data augmentation, the augmentation is based on synthesizing speech signals for the speech production characteristics of a specific talker, such as prosody and pitch. In [59], speech synthesis-based data augmentation was able to improve personalized speech enhancement performance compared to a generic system. In [60], zero-shot text-to-speech systems were used to augment data for training personalized speech enhancement systems, which outperformed generic systems. Differently, this paper investigates the significance of simulating transfer characteristics of own voice between hearable microphones for personalized data augmentation in order to train multi-channel OVR systems, addressing a gap in previous research.

While a direct comparison of text-to-speech-based data augmentation with the methods is outside the scope of this paper, it could be interesting to compare or even combine these techniques in future work. As a possible topic for further research, it would be interesting to investigate training personalized speech enhancement systems conditioned on auxiliary information about the target talker (see e.g., [61]) on a larger dataset with a sufficient amount of different talkers, such as Vibravox [62], and to compare between different approaches to personalization.

### 8.2 Comparison of OVR systems

The subjective listening test showed that OVR systems improved quality ratings over both noisy in-ear and noisy outer microphone signals. In addition, the OVR systems yield better performance than the considered baselines EBEN and MWF.

While personalized OVR was predicted to have a consistent benefit by instrumental metrics, this benefit was not consistent over all conditions in the results of the subjective listening test. The best performance is obtained when OVR systems are trained with generic data augmentation and personalized fine-tuning, although in some conditions there is no gain from personalized over generic fine-tuning, indicating that the benefits of personalization may be situation-dependent or limited. These benefits were only found between some conditions in the

subjective listening experiment, indicating either that the differences are too small to be noticed in other conditions or that the metrics incorrectly predicted this difference. Although we can only speculate about what caused the differences between talker 0 and talker 6, we note that the performance of generic systems is already quite good for talker 0 (*low predicted benefit*), while for talker 6 (*high predicted benefit*) worse results are achieved, leaving more room for improvement. Another possibility is that generic OVR systems perform better for talkers closer to the average talker (in terms of transfer characteristics or other speech production characteristics), requiring no personalization, and that for talkers that deviate from the average talker, personalization can yield improvement.

### 8.3 OVR quality prediction

In terms of quality prediction by instrumental metrics, the results have shown that not all metrics were able to predict subjective quality for OVR systems. Among the reference-based metrics, ESTOI, PESQ, and PEMO-Q PSM achieved the highest correlations. Among the reference-free metrics, LEAP, SCOREQ distance, and DNSMOS P808 achieved the highest correlations. In most cases, most metrics deviated from subjective ratings when predicting the quality of noisy or bandwidth-extended in-ear signals. This observation matches previously published research on subjective evaluation of bandwidth-extended [34] or body-conducted speech [23]. In [63] it was observed that both PESQ and POLQA were unable to correctly predict ranking of bandwidth extension algorithms. This was also observed in Figure 8 for the predictions of PESQ for the signals processed by EBEN. For in-ear microphone signals, in [33] the quality predictions obtained with PESQ were highly correlated with subjective ratings, which is consistent with the observations in this paper.

Other DNN-based metrics also exhibit deviations from subjective ratings of noisy or bandwidth-extended signals, which could be attributed to the fact that in-ear microphone signals are subject to degradations like individually different bandwidth limitations, body-produced noise, or time-varying changes in transfer characteristics. These degradations are not typical in standard single-channel speech enhancement evaluations as considered for the training of e.g., DNSMOS [43]. While WV-MOS was reported to achieve much higher correlation than PESQ or DNSMOS metrics for prediction of MOS for bandwidth extension in [64]<sup>3</sup>, the results in this paper do not reflect this. A possible explanation is the difference between signals band-limited by a lowpass filter and own voice signals recorded by an in-ear microphone. The high correlations of LEAP are somewhat surprising because the model was developed for listening effort predictions

<sup>3</sup> Note that this particular result is only reported in the extended preprint [64] and not in the conference paper [46].

and not for quality predictions. Comparing the results of eMoBi-Q and GPSM<sup>q</sup>, we observe that GPSM<sup>q</sup> achieves higher correlations and lower errors, which is likely due to the ability of GPSM<sup>q</sup> to account for non-linear moderate or heavy distortions. The ASR system underlying the phoneme classification had never seen strong band limitation or bandwidth extension algorithms during training. This may indicate the phoneme classification-based perception prediction generalizes well across different signal degradations and enhancement strategies, as suggested in [65] for different types of signal degradations.

## 9 Conclusion

In this paper, we have investigated personalized own voice reconstruction systems and evaluated their performance in terms of subjective quality. The systems were personalized during training with augmented data, or during fine-tuning with recorded data. Personalized systems were compared to their generic counterparts. While objective metrics predicted an increase in quality through personalization, the subjective evaluation only partly confirmed this improvement. In particular, most metrics failed to accurately predict the quality of bandwidth-limited noisy or bandwidth-extended in-ear microphone signals. This mismatch was observed when comparing objective predictions with median subjective ratings. Other metrics, such as the intrusive ESTOI and the non-intrusive LEAP, correlated well with subjective ratings. Both objective and subjective results demonstrate that the considered own voice reconstruction systems substantially increase own voice quality compared to noisy microphone signals and several baseline systems.

### Acknowledgments

We thank Fenja Hermann for conducting the listening experiment and Rainer Huber for helpful discussion and computing the LEAP and PEMO-Q PSM predictions. We also thank the study subjects for their participation in the listening experiment.

### Funding

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 work was partly funded by the German Ministry of Science and Education BMBF FK 16SV8811 and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 352015383 – SFB 1330 C1 and A1.

### Conflicts of interest

The authors declare no conflict of interest.

### Data availability statement

The research data associated with this article are available in Zenodo, under the references <https://doi.org/10.5281/zenodo.10844598> (recorded own voice signals) [66], <https://doi.org/10.5281/zenodo.11196866> (transfer function

measurements) [67], and <https://doi.org/10.5281/zenodo.15248719> (subjective ratings, objective predictions, and audio signals of the listening experiment stimuli) [68]. Selected audio examples are also available at [https://m-ohlenbusch.github.io/subjective\\_ovr\\_personalized/](https://m-ohlenbusch.github.io/subjective_ovr_personalized/).

### Informed consent

All subjects who participated in the recordings were informed about data collection and future data use, and gave informed consent.

### References

1. R.E. Bouserhal, T.H. Falk, J. Voix: Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones, in: Proceedings of the Meetings on Acoustics (ICA). Vol. 19. Montreal, QC, Canada, 2013.
2. S. Nordholm, A. Davis, P.C. Yong, H.H. Dam: Assistive listening headsets for high noise environments: protection and communication, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5753–5757.
3. R.E. Bouserhal, A. Bernier, J. Voix: An in-ear speech database in varying conditions of the audio-phonation loop. *Journal of the Acoustical Society of America* 145, 2 (2019) 1069–1077.
4. B. Gårdbæk, P. Kidmose: On the origin of cardiovascular sounds recorded from the ear. *IEEE Transactions on Biomedical Engineering* 72, 1 (2025) 210–216.
5. M.Ø. Hansen: Occlusion effects Part I and II. Ph.D. thesis, Department of Acoustic Technology, Technical University of Denmark, 1998.
6. T. Zurbrügg, A. Stirnemann, M. Kuster, H. Lissek: Investigations on the physical factors influencing the ear canal occlusion effect caused by hearing aids. *Acta Acustica United with Acustica* 100, 3 (2014) 527–536.
7. S. Stenfelt, S. Reinfeldt: A model of the occlusion effect with bone-conducted stimulation. *International Journal of Audiology* 46, 10 (2007) 595–608.
8. S. Vogl, M. Blau: Individualized prediction of the sound pressure at the eardrum for an earpiece with integrated receivers and microphones. *Journal of the Acoustical Society of America* 145, 2 (2019) 917–930.
9. S. Reinfeldt, P. Östli, B. Håkansson, S. Stenfelt: Hearing one’s own voice during phoneme vocalization – Transmission by air and bone conduction. *Journal of the Acoustical Society of America* 128, 2 (2010) 751–762.
10. H. Saint-Gaudens, H. Nélisse, F. Sgard, O. Doutres: Towards a practical methodology for assessment of the objective occlusion effect induced by earplugs. *Journal of the Acoustical Society of America* 151, 6 (2022) 4086–4100.
11. J. Richard, V. Zimpfer, C. Blondé-Weinmann, S. Roth: Change in transfer function between air and bone conduction microphones due to mouth opening variation. *Applied Acoustics* 228 (2025) 110293.
12. F. Denk, B. Kollmeier: The hearpiece database of individual transfer functions of an in-the-ear earpiece for hearing device research. *Acta Acustica* 5 (2021) 2.
13. K. Kondo, T. Fujita, K. Nakagawa: On equalization of bone conducted speech for improved speech quality, in: Proceedings of the International Symposium on Signal Processing and Information Technology, Vancouver, BC, Canada, 2006, pp. 426–431.
14. H.S. Shin, T. Fingscheidt, H.-G. Kang: A priori SNR estimation using air- and bone-conduction microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 11 (2015) 2015–2025.
15. R.E. Bouserhal, T.H. Falk, J. Voix: In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension. *Journal of the Acoustical Society of America* 141, 3 (2017) 1321–1331.
16. H. Wang, X. Zhang, D. Wang: Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022) 3134–3143.
17. J. Hauret, T. Joubaud, V. Zimpfer, É. Bavu: Configurable EBEN: Extreme Bandwidth Extension Network to enhance body-conducted speech capture. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023) 3499–3512.
18. M. Ohlenbusch, C. Rollwage, S. Doclo: Multi-microphone noise data augmentation for DNN-based own voice reconstruction for hearables in noisy environments, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, South Korea, 2024, pp. 416–420.
19. C. Li, F. Yang, J. Yang: A two-stage approach to quality restoration of bone-conducted speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024) 818–829.
20. C. Li, F. Yang, J. Yang: Restoration of bone-conducted speech with U-Net-like model and energy distance loss. *IEEE Signal Processing Letters* 31 (2024) 166–170.
21. A. Edraki, W.-Y. Chan, J. Jensen, D. Fogerty: Speaker adaptation for enhancement of bone-conducted speech, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, South Korea, 2024, pp. 10456–10460.
22. Y. Sui, M. Zhao, J. Xia, X. Jiang, S. Xia: TRAMBA: a hybrid transformer and Mamba architecture for practical audio and bone conduction speech super resolution and enhancement on mobile and wearable platforms, in: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. Vol. 8. New York, USA, 2024.
23. J. Richard, V. Zimpfer, S. Roth: Comparison of objective and subjective methods for evaluating speech quality and intelligibility recorded through bone conduction and in-ear microphones. *Applied Acoustics* 211 (2023) 109576.
24. K. Tesch, T. Gerkmann: Insights into deep non-linear filters for improved multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023) 563–575.
25. M. Ohlenbusch, C. Rollwage, S. Doclo: Speech-dependent data augmentation for own voice reconstruction with hearable microphones in noisy environments. *EURASIP Journal on Audio, Speech, and Music Processing* 2025 (2025) 32.
26. M. Ohlenbusch, C. Rollwage, S. Doclo: Low-complexity own voice reconstruction for hearables with an in-ear microphone, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025.
27. M. Ohlenbusch, C. Rollwage, S. Doclo: Modeling of speech-dependent own voice transfer characteristics for

- hearables with an in-ear microphone. *Acta Acustica* 8 (2024) 28.
28. F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J.-H. Bach, J. Wellmann, B. Kollmeier: A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research, in: *Proceedings of the AES International Conference on Headphone Technology*, San Francisco, USA, 2019.
  29. R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, G. Weber: Common voice: a massively-multilingual speech corpus, in: *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, 2020, pp. 4218–4222.
  30. H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh, R. Aichner: ICASSP 2023 deep noise suppression challenge. *IEEE Open Journal of Signal Processing* 5 (2024) 725–737.
  31. International Telecommunications Union (ITU): ITU-T P.862, perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, in: *International Telecommunications Union*, 2001.
  32. International Telecommunications Union (ITU): ITU-T P.863, perceptual objective listening quality prediction (POLQA), in: *International Telecommunications Union*, 2018.
  33. J.F. Santos, R. Bouserhal, J. Voix, T.H. Falk: Objective speech quality estimation of in-ear microphone speech, in: *Proceedings of the 5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)*, Berlin, Germany, 2016, pp. 69–73.
  34. P. Bauer, C. Guillaum e, W. Tirry, T. Fingscheidt: On speech quality assessment of artificial bandwidth extension, in: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 6082–6086.
  35. J. Jensen, C.H. Taal: An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 11 (2016) 2009–2022.
  36. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen: An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 7 (2011) 2125–2136.
  37. T. Biberger, J.-H. Fle ner, R. Huber, S.D. Ewert: An objective audio quality measure based on power and envelope power cues. *Journal of the Audio Engineering Society* 66, 7/8 (2018) 578–593.
  38. B. Eurich, S.D. Ewert, M. Dietz, T. Biberger: A computationally efficient model for combined assessment of monaural and binaural audio quality. *Journal of the Audio Engineering Society* 72, 9 (2024) 536–551.
  39. J.-H. Fle ner, T. Biberger, S.D. Ewert: Subjective and objective assessment of monaural and binaural aspects of audio quality. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 7 (2019) 1112–1125.
  40. T. Dau, D. P uschel, A. Kohlrausch: A quantitative model of the “effective” signal processing in the auditory system. I. Model structure. *Journal of the Acoustical Society of America* 99, 6 (1996) 3615–3622.
  41. R. Huber, B. Kollmeier: PEMO-Q – A new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 6 (2006) 1902–1911.
  42. A. Ragano, J. Skoglund, A. Hines: SCOREQ: speech quality assessment with contrastive regression, in: *Advances in Neural Information Processing Systems*. Vol. 37. 2024.
  43. C.K.A. Reddy, V. Gopal, R. Cutler: DNSMOS P.835: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, in: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 886–890.
  44. International Telecommunications Union (ITU): ITU-T P.835, subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, in: *International Telecommunications Union*, 2003.
  45. International Telecommunications Union (ITU): ITU-T P.808, subjective evaluation of speech quality with a crowdsourcing approach, in: *International Telecommunications Union*, 2018.
  46. P. Andreev, A. Alanov, O. Ivanov, D. Vetrov: HIFI++: a unified framework for bandwidth extension and speech enhancement, in: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023.
  47. R. Huber, M. Kr uger, B.T. Meyer: Single-ended prediction of listening effort using deep neural networks. *Hearing Research* 359 (2018) 40–49.
  48. R. Huber, A. Pusch, N. Moritz, J. Rannies, H. Schepker, B.T. Meyer: Objective assessment of a speech enhancement scheme with an automatic speech recognition-based system, in: *Proceedings of the ITG Conference on Speech Communication*, Oldenburg, Germany, 2018, pp. 86–90.
  49. M. Krueger, M. Schulte, T. Brand, I. Holube: Development of an adaptive scaling method for subjective listening effort. *Journal of the Acoustical Society of America* 141, 6 (2017) 4680–4693.
  50. International Telecommunications Union (ITU): ITU-R BS.1534-3, method for the subjective assessment of intermediate sound quality (MUSHRA), in: *International Telecommunications Union*, 2015.
  51. M. Schoeffler, S. Bartoschek, F.-R. St oter, M. Roess, S. Westphal, B. Edler, J. Herre: Web-MUSHRA – A comprehensive framework for web-based listening tests. *Journal of Open Research Software* 6, 1 (2018).
  52. S. Doclo, M. Moonen: GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing* 50, 9 (2002) 2230–2244.
  53. M. Souden, J. Chen, J. Benesty, S. Affes: Gaussian model-based multichannel speech presence probability. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 5 (2009) 1072–1077.
  54. S. Bagheri, D. Giacobello: Exploiting multi-channel speech presence probability in parametric multi-channel Wiener filter, in: *Proceedings of the Interspeech*, Graz, Austria, 2019, pp. 101–105.
  55. J. Rannies, M. Ohlenbusch, A. Volgenandt, T. Spitz, H. Baumgartner, C. Rollwage, V. Uslar, V. Weber: Analyse und algorithmische Optimierung von Ger uschkulissen und Sprachkommunikation im OP-Saal, in: *Proceedings of the German Annual Conference on Acoustics (DAGA)*, Hamburg, Germany, 2023, pp. 646–649.

56. R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2025. URL: <https://www.R-project.org/>.
57. J.G. Beerends, N.M.P. Neumann, E.L. van den Broek, A. Llagostera Casanovas, J.T. Menendez, C. Schmidmer, J. Berger: Subjective and objective assessment of full bandwidth speech quality. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019) 440–449.
58. L. He, H. Hou, S. Shi, X. Shuai, Z. Yan: Towards bone-conducted vibration speech enhancement on head-mounted wearables, in: *Proceedings of the Annual International Conference on Mobile Systems, Applications and Services*, New York, USA, 2023, pp. 14–27.
59. A. Kuznetsova, A. Sivaraman, M. Kim: The potential of neural speech synthesis-based data augmentation for personalized speech enhancement, in: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023.
60. J.-S. Bae, A. Kuznetsova, D. Manocha, J. Hershey, T. Kristjansson, M. Kim: Generative data augmentation challenge: zero-shot speech synthesis for personalized speech enhancement, in: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW): Generative Data Augmentation for Real-World Signal Processing Applications (GenDA 2025)*, Hyderabad, India, 2025.
61. K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, D. Yu: Neural target speech extraction: an overview. *IEEE Signal Processing Magazine* 40, 3 (2023) 8–29.
62. J. Hauret, M. Olivier, T. Joubaud, C. Langrenne, S. Poirée, V. Zimpfer, É. Bavu: Vibravox: a dataset of French speech captured with body-conduction audio sensors. *Speech Communication* 172 (2025) 103238.
63. H. Pulakka, V. Myllylä, A. Rämö, P. Alku: Speech quality evaluation of artificial bandwidth extension: comparing subjective judgments and instrumental predictions, in: *Proceedings of the Interspeech*, Dresden, Germany, 2015, pp. 2583–2587.
64. P. Andreev, A. Alanov, O. Ivanov, D. Vetrov: HIFI++: a unified framework for bandwidth extension and speech enhancement, 2023. [arXiv:2203.13086](https://arxiv.org/abs/2203.13086).
65. J. RENNIES, M. BERDAU, R. HUBER, H. BAUMGARTNER, S. WEIHE, T. BRAND: Real-time assessment of listening effort using non-intrusive binaural prediction models, in: *Proceedings of the DAS/DAGA 2025*, Copenhagen, Denmark, 2025, pp. 18–21.
66. M. Ohlenbusch, C. Rollwage, S. Doclo: German own voice recordings with hearable microphones. Zenodo, 2024. DOI: [10.5281/zenodo.10844599](https://doi.org/10.5281/zenodo.10844599). URL: <https://doi.org/10.5281/zenodo.10844599>.
67. M. Ohlenbusch, C. Rollwage, S. Doclo: Transfer function measurements for simulating environmental noise at hearable microphones. Zenodo, 2024. DOI: [10.5281/zenodo.11196867](https://doi.org/10.5281/zenodo.11196867). URL: <https://doi.org/10.5281/zenodo.11196867>.
68. M. Ohlenbusch, C. Rollwage, S. Doclo, J. RENNIES: Subjective ratings and objective metric predictions of generic and personalized own voice reconstruction systems. Zenodo, 2025. DOI: [10.5281/zenodo.15248719](https://doi.org/10.5281/zenodo.15248719). URL: <https://doi.org/10.5281/zenodo.15248719>.

**Cite this article as:** Ohlenbusch M. Rollwage C. Doclo S. & RENNIES J. 2026. Subjective quality evaluation of personalized own voice reconstruction systems. *Acta Acustica*, 10, 26. <https://doi.org/10.1051/aacus/2026021>.