

DNN-BASED OWN VOICE RECONSTRUCTION FOR
HEARABLES WITH AN IN-EAR MICROPHONE

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von
Herrn Mattes Ohlenbusch
geboren am 18. September 1995
in Oldenburg (Deutschland)

Mattes Ohlenbusch:

DNN-based own voice reconstruction for hearables with an in-ear microphone

ERSTGUTACHTER:

Prof. Dr. Simon Doclo

Carl von Ossietzky Universität Oldenburg, Germany

WEITERE GUTACHTER:

Prof. Dr. Jan Rennies-Hochmuth

Carl von Ossietzky Universität Oldenburg, Germany

Prof. Dr. Jesper Jensen

Aalborg University, Denmark

TAG DER DISPUTATION:

16. April 2026

ACKNOWLEDGMENTS

This thesis was written at the Signal Processing Group, Department of Medical Physics and Acoustics, at Carl von Ossietzky Universität Oldenburg, Germany. I would like to take the opportunity to thank the many people who helped me along the way.

First, I would like to thank Simon Doclo for his supervision, during which I have benefited immensely from his continuous guidance and support. I would also like to thank Christian Rollwage for many helpful discussions, ideas, and invaluable support, without which I would not have been able to complete this thesis. I am grateful to Jan Rennies-Hochmuth for reviewing my thesis, for serving in my defense committee, and for his cheerfulness and guidance throughout the years. I am thankful to Jesper Jensen for reviewing my thesis, for serving in my defense committee, and for showing much interest in my work.

Special thanks to the many current and former members of Fraunhofer HSA who have supported me in countless ways, in particular Paul, Ragini, Menno, Marc, Andi, Eray, Nils, Jordan, Malte, Anna, Robert, Till, AC, Bianca, Fenja, Steffen, and Christoffer. I also thank all participants in the recording sessions and listening tests.

I am deeply grateful to all my past and present colleagues in the UOL Signal Processing Group for their support and encouragement, as well as for many fun barbecues and conference trips. I am also very thankful to Miko, Marko, and Nils, who made my internship at Bose a very special, rewarding, and enjoyable experience. I would also like to thank Felix, Ragini, and Fabian for proofreading parts of this thesis.

Finally, I want to thank my parents, my brother, my friends, and my girlfriend for their continuous support and encouragement!

Oldenburg, May 15, 2026
Mattes Ohlenbusch

ABSTRACT

In recent years, hearable technology has advanced rapidly, leading to widespread daily use in challenging acoustic environments. As their popularity has grown, so has the demand for high-quality speech communication. Although hearables can capture the user’s own voice with outer microphones, recordings made in noisy conditions typically require processing to enhance speech quality, which can be challenging at high noise levels. Many modern hearables also include an in-ear microphone, which is more robust to environmental noise than the outer microphones because the device partially occludes the ear canal. However, in-ear own voice recordings exhibit characteristic distortions, such as low-frequency amplification and band-limitation, which vary strongly across individuals, change during speech production, and depend on device properties. These effects need to be taken into account when using an in-ear microphone for own voice capture.

The main objective of this thesis is to develop and evaluate causal deep neural network (DNN)-based own voice reconstruction (OVR) approaches that estimate clean broadband speech from noisy outer and in-ear microphone signals. Achieving this objective requires addressing several key challenges: understanding the unique distortions affecting in-ear own voice recordings, reducing the training data requirements of DNN-based OVR systems, meeting realistic computational complexity constraints, identifying suitable objective metrics for OVR performance that correlate well with subjective quality ratings, and investigating the benefits of personalizing OVR systems to individual talkers.

As a first contribution, we propose a phoneme-dependent model of the time-varying relationship between own voice signals recorded by an outer and an in-ear microphone, which we refer to as own voice transfer characteristics. Specifically, the model represents the own voice transfer characteristics as a set of linear time-invariant relative transfer functions, one for each phoneme. Experimental results on recorded own voice signals from 18 talkers demonstrate that the proposed (time-varying) phoneme-dependent model predicts in-ear own voice signals up to 50% more accurately than time-invariant models. While individual models yield lower prediction errors for matched talkers than talker-averaged models, talker-averaged models generalize better to unseen talkers.

As a second contribution, we propose data augmentation techniques for training multi-channel DNN-based OVR systems that jointly process the outer and in-ear microphone signals. The proposed augmentation technique, based on the phoneme-dependent own voice transfer characteristics model, enables the simulation of a large amount of in-ear own voice signals from a clean speech dataset, while requiring only a small amount of recorded own voice signals to identify the transfer characteristics

model. Experimental results for signal-to-noise ratios between -10 dB and 10 dB at the outer microphone show that the OVR system trained with phoneme-dependent individual augmentation followed by fine-tuning with recorded signals achieves the best performance, with an average PESQ improvement of 1.3 compared to the noisy outer microphone signal. This performance gain is maintained even when only a few minutes of recorded own voice signals per talker are available to identify the transfer characteristics model. In addition, to meet realistic computational constraints, we investigate low-complexity variants of the proposed DNN-based OVR system (down to 13 k parameters), and show that these variants outperform baseline OVR systems at comparable complexity.

As a third contribution, we investigate personalization of OVR systems to individual talkers using two approaches: training-based personalization and enrollment-based personalization. Results from a listening test show that generic (non-personalized) OVR systems substantially improve subjective quality compared to unprocessed noisy outer microphone signals with an average score improvement of 50 MUSHRA points, with personalization providing an additional benefit of up to 5 points for some talkers. A correlation analysis between objective metrics and subjective quality ratings indicates that the intrusive ESTOI metric and the non-intrusive LEAP metric are particularly suitable for assessing OVR performance. For the proposed enrollment-based personalization, an enrollment utterance of the talker recorded with the in-ear microphone is required. Experiments on the Vibravox dataset show that enrollment-based personalization is very effective in scenarios with competing talkers, achieving up to 10 dB SI-SDR improvement over unprocessed signals, and remains robust under dataset mismatch.

In summary, this thesis demonstrates that an OVR system combining an outer and an in-ear microphone can be trained with a small amount of recorded own voice signals by using the proposed phoneme-dependent own voice transfer characteristics models, enabling high-quality OVR for hearables in noisy environments. This is verified by objective metrics and the results of a subjective listening test.

ZUSAMMENFASSUNG

In den letzten Jahren hat sich die Hearable-Technologie schnell entwickelt, sodass Hearables heute in großem Umfang täglich in herausfordernden akustischen Umgebungen genutzt werden. Mit der steigenden Beliebtheit ist auch die Nachfrage nach hochqualitativer Sprachkommunikation gewachsen. Obwohl Hearables die Eigensprache der Nutzerin oder des Nutzers mit Außenmikrofonen aufnehmen können, erfordern Aufnahmen in lauten Umgebungen in der Regel eine Signalverarbeitung zur Verbesserung der Sprachqualität, die bei hohen Störgeräuschpegeln besonders schwierig ist. Viele moderne Hearables verfügen außerdem über ein In-Ohr-Mikrofon, das aufgrund der teilweisen Okklusion des Gehörgangs durch das Gerät gegenüber den Außenmikrofonen robuster gegenüber Umgebungsgeräuschen ist. Allerdings weisen mit dem In-Ohr-Mikrofon aufgezeichnete Eigensprachsignale charakteristische Verzerrungen auf, wie beispielsweise tieffrequente Verstärkung und Bandbegrenzung, die stark zwischen Personen variieren, sich während der Sprachproduktion verändern und von den Geräteeigenschaften abhängen. Diese Effekte müssen bei der Nutzung eines In-Ohr-Mikrofons zur Aufnahme der Eigensprache berücksichtigt werden.

Das Hauptziel dieser Dissertation ist die Entwicklung und Bewertung kausaler, auf tiefen neuronalen Netzen (engl. *deep neural networks*, DNNs) basierenden Verfahren zur Rekonstruktion der Eigensprache (engl. *own voice reconstruction*, OVR), die aus störgeräuschbehafteten Signalen eines Außen- und eines In-Ohr-Mikrofons störgeräuschbefreite breitbandige Sprache schätzen. Um dieses Ziel zu erreichen, müssen mehrere zentrale Herausforderungen adressiert werden: das Verständnis der spezifischen Verzerrungen in mit In-Ohr-Mikrofonen aufgezeichneten Eigensprachsignalen, die Reduktion des Trainingsdatenbedarfs DNN-basierter OVR-Systeme, die Einhaltung realistischer Anforderungen an die Rechenkomplexität, die Identifikation geeigneter objektiver Kenngrößen zur Bewertung der OVR-Performance mit guter Korrelation zu subjektiven Qualitätsurteilen sowie die Untersuchung des Nutzens einer Personalisierung von OVR-Systemen auf individuelle Sprecherinnen und Sprecher.

Als ersten Beitrag schlagen wir ein phonemabhängiges Modell der zeitvarianten Beziehung zwischen Eigensprachsignalen vor, die mit einem Außen- und einem In-Ohr-Mikrofon aufgezeichnet werden, welches wir als Übertragungscharakteristika der Eigensprache (engl. *own voice transfer characteristics*) bezeichnen. Konkret beschreibt das Modell die Übertragungscharakteristika der Eigensprache als eine Menge linear zeitinvarianter relativer Übertragungsfunktionen, von denen jede einem Phonem zugeordnet ist. Experimentelle Ergebnisse mit aufgezeichneten Eigensprachsignalen von 18 Sprecherinnen und Sprechern zeigen, dass das vorgeschlagene (zeitvariante) phonemabhängige Modell In-Ohr-Eigensprachsignale bis zu 50% ge-

nauer vorhersagt als zeitinvariante Modelle. Während individuelle Modelle für die jeweils zugehörigen Sprecherinnen und Sprecher geringere Vorhersagefehler liefern als über Sprecher gemittelte Modelle, generalisieren über Sprecher gemittelte Modelle besser auf unbekannte Sprecherinnen und Sprecher.

Als zweiten Beitrag schlagen wir Datenaugmentierungstechniken zum Training mehrkanaliger, DNN-basierter OVR-Systeme vor, die die Signale des Außen- und des In-Ohr-Mikrofons gemeinsam verarbeiten. Die vorgeschlagene Augmentierungstechnik, die auf dem phonemabhängigen Modell der Übertragungscharakteristika der Eigensprache basiert, ermöglicht die Simulation einer großen Anzahl von In-Ohr-Eigensprachsignalen aus einem Datensatz in Ruhe aufgezeichneter Sprache, während zur Identifikation des Modells der Übertragungscharakteristika nur eine geringe Menge aufgezeichneter Eigensprachsignale benötigt wird. Experimentelle Ergebnisse für Signal-Störgeräusch-Abstände zwischen -10 dB und 10 dB am Außenmikrofon zeigen, dass das mit phonemabhängiger individueller Augmentierung trainierte und anschließend mit aufgezeichneten Signalen feinjustierte OVR-System die beste Leistung erzielt und im Mittel eine PESQ-Verbesserung von 1.3 gegenüber dem störgeräuschbehafteten Außenmikrofonsignal erreicht. Selbst wenn zur Identifikation des Modells der Übertragungscharakteristika pro Sprecherin oder Sprecher nur wenige Minuten aufgezeichneter Eigensprache zur Verfügung stehen, bleibt dieser Leistungsgewinn erhalten.

Als dritten Beitrag untersuchen wir die Personalisierung von OVR-Systemen auf individuelle Sprecherinnen und Sprecher mit zwei Ansätzen: trainingsbasierte Personalisierung und anmeldungsbasierte (engl. *enrollment-based*) Personalisierung. Die Ergebnisse eines Hörtests zeigen, dass generische (nicht-personalisierte) OVR-Systeme die subjektive Sprachqualität gegenüber unverarbeiteten, störgeräuschbehafteten Außenmikrofonsignalen deutlich verbessern und im Mittel einen Gewinn von 50 MUSHRA-Punkten erzielen, wobei Personalisierung für einige Sprecherinnen und Sprecher einen zusätzlichen Gewinn von bis zu 5 Punkten erzielt. Eine Korrelationsanalyse zwischen objektiven Metriken und subjektiven Qualitätsbewertungen zeigt, dass die intrusive Metrik ESTOI und die nicht-intrusive Metrik LEAP sich besonders gut zur Bewertung der OVR-Performance eignen. Für die vorgeschlagene anmeldungsbasierte Personalisierung ist eine Anmeldungs-Äußerung der Sprecherin oder des Sprechers erforderlich, die mit dem In-Ohr-Mikrofon aufgezeichnet wird. Experimente auf dem Vibravox-Datensatz zeigen, dass anmeldungsbasierte Personalisierung in Szenarien mit konkurrierenden Sprecherinnen und Sprechern sehr effektiv ist, Verbesserungen der SI-SDR von bis zu 10 dB gegenüber unverarbeiteten Signalen erzielt und auch beim Testen auf einem anderen Datensatz robust bleibt.

Zusammenfassend zeigt diese Dissertation, dass ein OVR-System, das ein Außen- und ein In-Ohr-Mikrofon kombiniert, mithilfe der vorgeschlagenen phonemabhängigen Modelle der Übertragungscharakteristika der Eigensprache mit einer kleinen Menge aufgezeichneter Eigensprachsignale trainiert werden kann und so eine hochqualitative OVR für Hearables in lauten Umgebungen ermöglicht. Dies wird durch objektive Metriken und die Ergebnisse eines subjektiven Hörtests bestätigt.

LIST OF ABBREVIATIONS

AH	artificial head
AOC	active occlusion cancellation
AS-SE	auxiliary-sensor speech enhancement
DA	data augmentation
DNN	deep neural network
DNSMOS	deep noise suppression mean opinion score
EBEN	extreme bandwidth extension network
eMoBi-Q	efficient model for binaural audio quality
ERB	equivalent rectangular bandwidth
ESTOI	extended short-time objective intelligibility
FIR	finite impulse response
FT	fine-tuning
FT-JNF	frequency-and-time joint non-linear filter
GAN	generative adversarial network
GCBFSNet	group communication binaural filter-and-sum network
GPSM ^a	generalized power spectrum model for quality
GRU	gated recurrent unit
IM	in-ear microphone
LEAP	listening effort prediction from acoustic parameters
LSD	log-spectral distance
LSTM	long short-term memory
MACs	multiply-accumulate operations
MCD	mel-cepstral distance

MUSHRA	multiple stimuli with hidden reference and anchor
MWF	multi-channel Wiener filter
NLMS	normalized least mean squares
OM	outer microphone
OVR	own voice reconstruction
PAS-SE	personalized auxiliary-sensor speech enhancement
PESQ	perceptual evaluation of speech quality
PSE	personalized speech enhancement
PSM	perceptual similarity measure
RF	real-time factor
RMSE	root mean squared error
RTF	relative transfer function
SCOREQ	speech contrastive regression for quality
SE	speech enhancement
SI-SDR	scale-invariant signal-to-distortion ratio
SNR	signal-to-noise ratio
SPL	sound pressure level
STFT	short-time Fourier transform
STOI	short-time objective intelligibility
TSE	target speaker extraction
VAD	voice activity detection
WOLA	weighted overlap-add

CONTENTS

1	Introduction	1
1.1	Acoustic scenario	2
1.2	State-of-the-art OVR approaches	4
1.3	Thesis outline and main contributions	14
	References	17
2	Modeling of speech-dependent own voice transfer characteristics for hearables with an in-ear microphone	27
2.1	Introduction	29
2.2	Signal model	31
2.3	Modeling of own voice transfer characteristics	32
2.4	Experimental evaluation	37
2.5	Discussion	46
2.6	Conclusion	48
	References	49
3	Speech-dependent data augmentation for own voice reconstruction with hearable microphones in noisy environments	53
3.1	Introduction	55
3.2	Own voice transfer characteristic models	57
3.3	Data augmentation techniques	60
3.4	Experimental setup	62
3.5	Experimental results	68
3.6	Discussion	74
3.7	Conclusion	75
	References	77
4	Low-complexity own voice reconstruction for hearables with an in-ear microphone	83
4.1	Introduction	85
4.2	Signal model	86
4.3	Own voice reconstruction system	86
4.4	Phoneme-dependent own voice augmentation	87
4.5	Experimental setup	88
4.6	Results	90
4.7	Conclusion	92
	References	93
5	Subjective quality evaluation of personalized own voice reconstruction systems	97
5.1	Introduction	99
5.2	Own voice reconstruction	100
5.3	OVr system training setup	104

5.4	Objective quality prediction metrics	105
5.5	Results of instrumental assessment	107
5.6	Listening experiment	109
5.7	Results of listening experiment	112
5.8	Discussion	119
5.9	Conclusion	121
	References	122
6	PAS-SE: Personalized auxiliary-sensor speech enhancement for voice pickup in hearables	129
6.1	Introduction	131
6.2	Signal model	132
6.3	System architecture	132
6.4	Evaluation details	133
6.5	Results	135
6.6	Conclusion	139
	References	140
7	Discussion	143
7.1	Chapter-by-chapter discussion	143
7.2	General discussion	152
	References	153
8	Conclusion	159
A	Multi-microphone noise data augmentation for DNN-based own voice reconstruction for hearables in noisy environments	161
A.1	Introduction	163
A.2	Signal model	164
A.3	Noise data augmentation	164
A.4	Experimental setup	166
A.5	Results	168
A.6	Conclusion	170
	References	170
	List of Publications	173

INTRODUCTION

In recent years, headphone technology has rapidly advanced, leading to widespread daily use. As headphones become more prevalent, they are increasingly used in various acoustic scenarios, e.g., in public transport, during travel, or at work. Such acoustic scenarios often include unwanted environmental noise sources or interfering talkers. Due to these unwanted acoustic sources, there is an increased demand for headphones to be able to change or block out parts of their acoustic environment. Prominent examples include headphones with active noise reduction to reduce environmental noise [1], acoustic transparency to enable the user to be aware of their surroundings [2], and spatial audio to alter the perceived acoustic environment [3, 4]. The acoustic environment of a headphone user can be modified more effectively if the headphone is able to capture it, for example by using microphones or other sensors included in the device. In this work, we refer to headphones equipped with loudspeakers, microphones, and other sensors as hearables. As their capabilities have increased, the demand for voice communication with hearables has also grown. Because hearables include one or more microphones, they can capture the user's own voice, which can be transmitted to human communication partners or voice assistants [5–8]. However, in challenging acoustic scenarios, the recorded own voice can be hard to understand and lacks quality due to also recording environmental noise and interfering talkers, limiting its suitability for direct transmission. Preprocessing is therefore required to improve intelligibility and quality. In recent years, several algorithms for noise reduction and speaker separation have been proposed that use one or more outer microphones on the hearable [9–15]. To improve their performance, it has been suggested to also use auxiliary sensors included in hearables.

For example, many hearables include an accelerometer to track head movements or vital signs, to enable gesture control, or to detect device insertion and removal [16], but they can also be used for own voice pickup [17, 18]. In addition, many hearables contain an in-ear microphone [19, 20], which can be beneficial for active noise reduction [1, 21, 22], active occlusion control [23], individualized sound pressure prediction and equalization [24, 25], and voice control [7]. Because in-ear microphones are partially shielded from environmental noise due to the hearable occluding the ear canal, they offer unique advantages for own voice pickup in noisy acoustic environments. Nevertheless, in-ear microphones also pose challenges: The own voice recorded by an in-ear microphone is mostly conducted through bone and cartilage

of the user’s head. Own voice recorded inside the occluded ear is predominantly body-conducted and affected by the occlusion effect, with frequency content below approximately 1 kHz being amplified, and frequency content above approximately 2 kHz being band-limited [26]. These distortion effects strongly vary across individuals [25, 27–29], and change during speech production [30–32]. They also depend on device properties, such as earmold fit and insertion depth [33, 34].

Aiming to estimate clean broadband speech from distorted own voice recordings with in-ear microphones or other auxiliary sensors, several own voice reconstruction (OVR) algorithms have been proposed. The objective of an OVR algorithm is not only to perform bandwidth extension and equalization, but also to reduce environmental noise leaking through the hearable and body-produced noise, possibly personalized to the individual wearing the hearable. For OVR, both classical signal processing approaches [35, 36] and machine learning approaches based on deep neural networks (DNNs) [37–39] have been proposed. It should be realized that DNN-based approaches require substantial training data, in particular own voice recordings from multiple talkers wearing the hearable device, in order to capture individual variation. In addition, several existing OVR systems have high computational complexity or even require non-causal processing (e.g., [40, 41]), making these approaches unsuitable for on-device implementations.

The main objective of this thesis is to develop and evaluate OVR systems for hearables with an in-ear microphone, focusing on causal systems with moderate algorithmic latency. To achieve this, we address the following key challenges:

- Modeling the own voice transfer characteristics between the outer face of the hearable and the in-ear microphone
- Reducing training data requirements and computational complexity of DNN-based OVR systems
- Personalization of OVR systems
- Identifying suitable objective metrics for evaluation

In the remainder of this chapter, we introduce the considered acoustic scenario in Section 1.1, we provide an overview of the state of the art in OVR systems in Section 1.2, and we outline the main contributions and the structure of this thesis in Section 1.3.

1.1 Acoustic scenario

In this thesis, we consider a hearable device equipped with an outer microphone at the outer face of the hearable and an in-ear microphone inside the (partly) occluded ear canal. The hearable is worn by a person, referred to as talker, in a noisy acoustic environment. We refer to speech by the talker as own voice, which is picked up by both the outer and the in-ear microphone. When recording own voice at the outer microphone in noisy environments, environmental noise and interfering talkers are

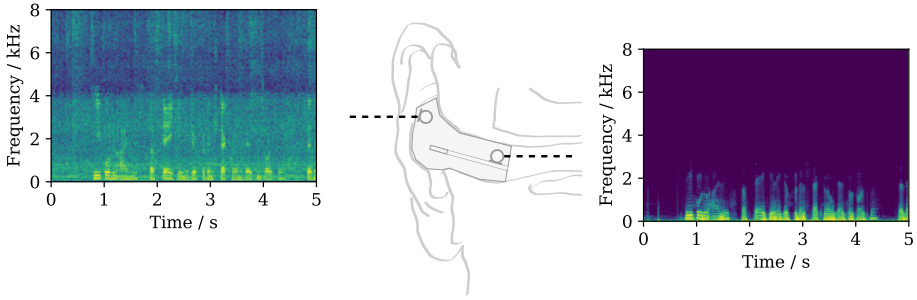


Figure 1.1: Spectrograms of noisy own voice signals recorded at the outer microphone (left) and at the in-ear microphone (right). While the outer microphone records broadband own voice, it also records environmental noise. The in-ear microphone mostly picks up body-conducted own voice, which is amplified below 1 kHz and band-limited at 2 kHz compared to own voice at the outer microphone. Environmental noise at the in-ear microphone is significantly attenuated.

also recorded. When recording own voice at the in-ear microphone, environmental noise and interfering talkers are significantly attenuated due to the hearable device (at least partly) occluding the ear canal. We refer to this residual environmental noise recorded at the in-ear microphone as noise leakage. The amount of noise leakage depends on the device and on how well the device fits the individual ear of the talker. In practice, both microphones also record sensor noise at a very low amplitude. In addition, the in-ear microphone also records body-produced noise (e.g., respiratory and cardiovascular sounds [42]) at a low amplitude. Figure 1.1 shows an example of noisy own voice signals recorded at the outer microphone and the in-ear microphone. Since neither the quality of the outer microphone nor the quality of the in-ear microphone is sufficient, several OVR approaches have been proposed that aim at estimating clean broadband speech from the outer microphone signal and/or the in-ear microphone signal. In the remainder of this thesis, we will assume this acoustic scenario and goal unless stated otherwise. Detailed mathematical signal models are introduced in the following chapters.

Unless stated otherwise, we only consider one ear (i.e., a single hearable). In Chapters 2 to 5, we only consider the closed-fit Hearpiece device [19] (i.e., without a vent), while in Chapter 6 we consider both the Hearpiece and a closed-fit soft-foam prototype device [43]. In Section 1.2, we provide an overview of OVR approaches that consider not only in-ear microphones but also vibration sensors and accelerometers, because of their similarity to in-ear microphones. We refer to in-ear microphones, vibration sensors, and accelerometers used for OVR as auxiliary sensors. It is important to note that while approaches using vibration sensors and accelerometers often do not consider environmental noise (since they record structural vibration of bone and cartilage), in-ear microphones also record the air-conducted environmental noise leakage transmitted through the hearable device.

1.2 State-of-the-art OVR approaches

Previous research has investigated in-ear microphones and other auxiliary sensors such as vibration sensors and accelerometers for own voice pickup. As shown in Figure 1.2, we categorize OVR systems into single-channel approaches (using only the in-ear microphone or auxiliary sensor) and multi-channel approaches (using both an outer microphone and an in-ear microphone or auxiliary sensor).

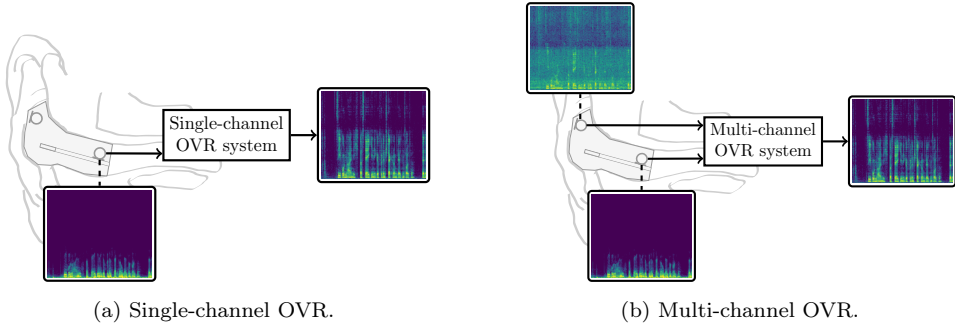


Figure 1.2: Single- and multi-channel own voice reconstruction, aiming at combining bandwidth extension, equalization, and noise reduction. Single-channel OVR processes only the in-ear own voice signal, while multi-channel OVR processes outer and in-ear own voice signals.

Section 1.2.1 provides an overview of classical signal processing approaches for OVR. Recently, deep learning-based approaches have demonstrated promising reconstruction performance compared to classical approaches. Section 1.2.2 gives an overview of single-channel DNN-based OVR approaches using only one in-ear microphone or auxiliary sensor. Aiming to improve the performance of approaches processing only the in-ear microphone signal, several multi-channel DNN-based OVR approaches that use both the in-ear and outer microphone signals have been proposed. Using multiple microphones allows these systems to overcome the limitations of band-limited signals and environmental noise, resulting in higher speech quality than single-channel approaches. Section 1.2.3 gives an overview of multi-channel DNN-based OVR approaches, which constitute the main focus of this thesis. Although in the remainder of this thesis we will only consider in-ear microphones, we also discuss approaches using other auxiliary sensors in this section because of their similarity to in-ear microphones regarding noise robustness and bandwidth limitations. We will also discuss some approaches that use a close-talk or a remote microphone, such as in front of the talker, which will also record broadband own voice similar to own voice recorded at an outer microphone. However, it should be realized that close-talk microphones are usually not available in practical applications.

1.2.1 *Classical approaches*

Following [44], we group classical signal processing approaches to reconstruct own voice from an in-ear microphone or another auxiliary sensor into equalization approaches, analysis-and-synthesis approaches, and probabilistic approaches. In [27], equalization filters were applied to in-ear microphone own voice signals to approximate the frequency response of own voice signals recorded at a close-talk microphone. Experimental results showed that talker-specific equalization filters achieved the best results in terms of objective metrics and subjective ratings. Although equalization can compensate for the low-frequency amplification observed in in-ear own voice signals, linear filtering **cannot restore missing high-frequency speech content** and may amplify sensor and body-produced noise. In [45], an approach based on linear prediction analysis and synthesis was proposed for body-conducted speech. Experiments with two female and two male talkers revealed large performance differences across talkers. Other works explored probabilistic approaches informed by an auxiliary sensor. For example, the approach proposed in [46] used an in-ear microphone signal to improve parameter estimation in outer microphone noise reduction. Similarly, in [47] a vibration sensor was used for voice activity detection (VAD) to improve the noise power estimation at the outer microphone.

Whereas in [46, 47] the auxiliary sensor signal only supported parameter estimation for single-channel noise reduction, approaches have been proposed where both the outer microphone signal and the auxiliary sensor signal are filtered. In [35], the auxiliary sensor signal was first equalized before being combined with the outer microphone signal in a probabilistic noise reduction scheme. Alternatively, in [36] an adaptive filter was used to remove noise leakage from an outer to an in-ear microphone. After removing residual noise from the in-ear microphone, a non-linear bandwidth extension approach was used to reconstruct high-frequency speech content from the denoised in-ear own voice signals. The approach considerably improved subjective quality compared to unprocessed or denoised in-ear own voice signals. Another approach for combining an outer microphone and a vibration sensor in smart glasses was proposed in [48], mainly focusing on wind noise reduction. In this approach based on dictionary learning, the noise components at each microphone were modeled independently of each other, while the own voice components at both microphones were modeled jointly. It is assumed that own voice at the outer and the vibration sensor is correlated, but wind noise at the outer microphone is uncorrelated with body-produced noise at the vibration sensor, leading to improved own voice estimation from using the vibration sensor compared to only using the outer microphone. Using both sensors led to considerable improvements over using either sensor alone in an evaluation with four recorded talkers. Similarly, [49] proposed combining a close-talk microphone and a throat vibration sensor using non-negative matrix factorization, switching processing between voice and unvoiced segments.

In contrast, classical signal processing approaches for own voice enhancement that do not consider auxiliary sensors have also been proposed. These approaches focus primarily on larger microphone arrays, a combination of hearing aids and remote microphones [50], or smart glasses equipped with multiple microphones [51], with-

out considering in-ear microphones or auxiliary sensors. In [50], binaural hearing aid processing was split between a fixed low-delay beamformer for own voice and an adaptive beamformer with higher processing delay for external sounds. This split was designed to minimize delay in own voice processing, which hearing device users are sensitive to, while allowing higher processing delay for external sounds which relaxes algorithmic constraints. The fixed distortionless beamformer was designed to isolate own voice (assumed to originate between the two hearing aids). For external sounds, the adaptive filter using microphones of both hearing aids and additional remote microphones, using information from the own voice processing path. In [51], a combination of adaptive and switching beamforming was proposed to isolate own voice from noisy signals recorded using smart glasses with multiple microphones. In particular, the approach considers a microphone between the lenses, which may be occluded. In this approach, the adaptive beamformer estimates noisy and noise covariance matrices informed by an oracle VAD, while oracle occlusion detection is used to switch between different sets of beamformer coefficients for the occluded and the unoccluded case. Both approaches in [50, 51] consider microphone arrays larger than what is practical for a single in-the-ear hearable. Both approaches also rely on VAD, which may be prone to errors in environments with **heavy noise**. As suggested in [50], it may be beneficial for these approaches to also consider auxiliary sensors for this task. However, auxiliary sensors may be difficult to integrate into classical algorithms due to the unique properties of own voice recorded with auxiliary sensors. For this reason, many current OVR approaches are based on DNNs.

1.2.2 *Single-channel DNN-based approaches*

DNN-based approaches have achieved large improvements in research areas that are closely related to OVR, for example in bandwidth extension [52–56] and speech enhancement [57–69]. Consequently, DNN-based approaches have also become more popular for OVR. Due to their non-linear processing capabilities, DNN-based approaches are able to reconstruct high-frequency speech content missing in in-ear own voice signals, combine in-ear and outer microphone signals non-linearly, or both. While single-channel speech enhancement approaches could also be used for OVR, their performance is limited in scenarios in low signal-to-noise ratios (SNRs) [70]. Approaches to bandwidth extension can be very sensitive to the filter response limiting the bandwidth of the input signal [55]. This could be a problem for OVR using an in-ear microphone, where individual differences between talkers can lead to differences in bandwidth, and equalization and noise reduction also need to be performed. Due to these limitations of single-channel speech enhancement and bandwidth extension when applied to OVR, the overview in this section focuses on single-channel DNN-based approaches designed specifically for in-ear microphone or other auxiliary sensor signals. Table 1.1 gives an overview of single-channel DNN-based OVR approaches, in terms of training and test data, use of additional fine-tuning with recorded data, environmental noise handling, and model size in terms of number of DNN parameters (if reported).

Table 1.1: Overview of single-channel DNN-based OVR approaches. IM noise: In-ear microphone (environmental) noise, FT: Fine-tuning with recorded data.

1 st	Author	Year	Ref.	Training data	FT	Test data	IM noise	Parameters
	Liu	2018	[71]	recorded ¹	-	recorded ¹	-	-
	Park	2019	[72]	recorded	-	recorded	-	-
	Nguyen	2020	[73]	recorded	-	recorded	-	-
	Ohlenbusch	2022	[74]	simulated	✓	recorded	-	10.2 M
	Zheng	2022	[40]	recorded	-	recorded	-	0.52 M ²
	Hauret	2023	[39, 75]	simulated	-	simulated	-	1.9 M
	Li	2023	[76]	recorded	(✓) ³	recorded	✓ ³	≥ 4.5 k
	Li	2024	[77]	recorded	-	recorded	-	-
	Edraki	2024	[78]	simulated	✓	recorded	-	-
	Sui	2024	[18]	simulated ⁴	✓	recorded	✓	5.2 M
	Li	2024	[79]	recorded	-	recorded	-	12.94 M ²
	Hauret	2025	[80]	recorded	✓	recorded	(✓) ⁵	96.2 M

¹ Recordings of a single talker were used.

² Parameter count according to [81].

³ Noise recordings of a single talker were used for a separate fine-tuning experiment.

⁴ DNN first trained for standard bandwidth extension, and fine-tuned for OVR.

⁵ Demo paper only contains results without IM noise; an example with noise is shown in a video.

Some DNN-based approaches extend the classical approaches: In [72, 73] frequency-domain equalization filtering is supported by a DNN, either for estimation of filter coefficients using a multi-layer feedforward network [72] or for estimation of line spectral frequencies used in filter computation by a long short-term memory (LSTM)-based network [73]. Both approaches share the limitations of classical equalization approaches, which are the **inability to restore missing high-frequency speech content** and the potential amplification of sensor and body-produced noise.

In contrast, many other approaches use DNNs directly to reconstruct high frequency content missing from auxiliary sensor own voice signals [18, 39, 40, 71, 74–80]. Processing is performed either in the time domain [40, 74, 79] or in the time-frequency domain [39, 71, 75–78]. Different auxiliary sensors are considered: vibration sensors [40, 76, 80], accelerometers [18, 77, 79], and in-ear microphones [39, 74, 75].

Many OVR approaches using auxiliary sensors, such as accelerometers or vibration sensors, **assume that the sensor does not pick up environmental noise**, since it primarily records structural vibration from the body [39, 71, 74, 75, 77–79]. However, this assumption does not hold for OVR approaches using in-ear microphones as auxiliary sensors, where environmental noise leakage can be amplified and lead to audible distortions and therefore needs to be considered.

For DNN-based approaches, the **availability of realistic training data** can also be a limiting factor. For instance, in [71] an autoencoder neural network was used to reconstruct own voice from auxiliary sensor signals, but was only trained and tested on recorded signals of the same single talker. Due to individual differences in own voice production and transmission to the in-ear microphone, this approach would likely not generalize to other talkers. Recording own voice signals with an auxiliary sensor for training an OVR system can be costly, since recordings from many utterances of multiple talkers may be required. Hence, previous approaches have addressed the simulation of auxiliary sensor own voice signals for training OVRs systems [18, 39, 74, 75, 78, 82]: In our previous work [74], we trained an OVR system with simulated own voice signals. The system used a UNet architecture based on a speech bandwidth extension approach [55]. The simulation used multiple estimated relative transfer functions (RTFs) between an outer and an in-ear microphone (assuming a linear time-invariant transfer function) and additive body-produced noise. Different simulation conditions with a single or multiple RTFs per talker (estimated on different speech segments) from one or multiple talkers were compared. It was found in an experimental evaluation with recorded in-ear own voice signals from 14 talkers that OVR performance increased when the number of RTFs per talker or the number of talkers was increased. Fine-tuning with recorded own voice signals considerably improved the trained system’s performance compared to training with simulated signals only. Unlike [72, 73], this approach generated high-frequency content, and unlike [71], it was evaluated on different talkers than those used for DNN training, demonstrating generalization to unseen talkers. In [39, 75], a generative adversarial network (GAN) scheme for OVR was proposed. The system was trained using own voice signals simulated similarly to [74], but unfortunately was also only evaluated using simulated signals. This presents a limitation, since evaluation with simulated signals matches the simulated training and the simulation might not reflect realistic recorded signals. To investigate the generalization of systems trained using simulated data to realistic conditions, recorded data should be used for testing, as considered in [74] and also in [18, 78].

Despite using personalization to improve performance, recent OVR approaches remain limited by strong individual differences in own voice transmission, which hinder **generalization to talkers not seen during training**. In [78], it was proposed to extend the bandwidth of own voice recorded at an in-ear microphone using a personalized approach. Similar to [74], the DNN was first trained with simulated signals, and then fine-tuned with recorded signals. While the approach was able to consistently improve speech quality for talkers in the training data, generalization to unseen talkers was reported to be poor. This presents a limitation for practical applications in which a sufficient amount of recorded signals of the target talker is not available. In [18], an OVR system was first trained for standard bandwidth extension and then fine-tuned with recorded signals. The system consisted of a hybrid transformer-Mamba architecture. The same OVR system architecture was applied to two sensors separately, comparing the approach for a vibration sensor or for an accelerometer. The performance was better when using a vibration sensor than when using an accelerometer. While personalized fine-tuning led to improvement compared to generic (non-personalized) system, poor generalization to talkers not

included in the training data was reported, which limits the approach to application in which a sufficient amount of recorded signals of the target talker are available. In [76], a DNN based on the UNet architecture [83] was trained and evaluated on recorded own voice signals. A fine-tuning step was employed to improve the robustness against environmental noise recorded at a vibration sensor. The evaluation of noise robustness was carried out by fine-tuning and testing with recorded noise signals of a single talker mixed with recorded own voice signals of multiple talkers. It is unclear whether this noise fine-tuning method is able to generalize to OVR with an in-ear microphone, where environmental noise leakage is subject to individual differences. It should be noted that the system proposed in [76] was reported to have only 4.5k trainable parameters. In [79], another UNet-like architecture for OVR with an accelerometer was proposed (consisting of 12.94 M parameters) and compared to other approaches proposed in [37, 40, 75, 76]. Importantly, the authors reported that the approach in [76] performed poorly in an experimental evaluation.

For practical applications of OVR, approaches need to meet **computational complexity requirements** of hardware included in hearables. In addition, **causal processing is required** for real-time communication applications. In [40], a non-causal transformer-based architecture was proposed for reconstructing own voice from vibration sensor signals. The approach employed both an equalization module and a generation module to generate the missing high-frequency speech content. The system was evaluated by training on own voice signals of one talker at a time (resulting in a personalized system), and testing on different utterances of the same talker. Due to the non-causal processing, this approach is not suitable for real-time applications. In terms of computational complexity, not all approaches report number of trainable parameters or number of required operations. Some approaches consist of several million parameters, e.g., [18, 74, 79, 80] (10.2 M, 5.2 M, 12.94 M, and 96.2 M parameters, respectively). In particular, in [80], a real-time OVR system for a throat vibration sensor was proposed based on a fine-tuned neural audio codec, reportedly robust to environmental noise without including noise in the training. However, the system consisted of 96.2 M parameters. For running on an embedded device, the complexity of the system would likely need to be reduced.

Evidently, these previously discussed single-channel DNN-based OVR approaches exhibit different limitations when applied to OVR using an in-ear microphone. First, approaches based on equalization filtering [72, 73] are inherently limited due to their inability to restore high-frequency speech content from band-limited signals. Second, many approaches do not consider environmental noise [39, 71–75, 77–79]. Third, while many approaches use simulated signals as training data, the majority of them relies on fine-tuning with recorded signals to improve performance. This indicates that the simulation procedures are not able to realistically model in-ear own voice signals. Simulation could likely be improved by considering factors that influence own voice at an in-ear microphone, such as individual differences or the phonemes being uttered. Fourth, while fine-tuning can be used to improve performance of a generic system, or to perform training-based personalization by fine-tuning recorded signals of a single talker, the role of the amount of recorded data is unclear. In

addition, generalization to unseen talkers is desirable, since recording new in-ear or auxiliary sensor own voice signals for fine-tuning further limits applicability. Similarly, the influence of recorded noise for training or fine-tuning and whether individual differences need to be accounted for have not been investigated. Finally, practical applications impose constraints on computational complexity and require causal processing. These limitations warrant further investigation into OVR using an in-ear microphone, in terms of realistic own voice and environmental noise data augmentation, reduction of computational complexity, and personalization of OVR processing to account for individual differences.

1.2.3 *Multi-channel DNN-based approaches*

As previously mentioned, several multi-channel DNN-based OVR approaches that use both the in-ear and outer microphone signals have been proposed to achieve better OVR performance compared to approaches processing only the in-ear microphone signal by overcoming the limitations of band-limited signals and environmental noise. Table 1.2 presents an overview of recent multi-channel DNN-based approaches. Many of the proposed approaches have been compared to single-channel approaches using only the outer microphone, e.g., [17, 37, 38, 84–86] or to approaches using only the auxiliary sensor, e.g., [37, 38, 84], demonstrating improvements from considering both channels as input signals. While some of the approaches consider time-domain processing [17, 37, 84, 87, 88], other approaches perform time-frequency domain processing [8, 38, 41, 81, 82, 85, 86, 89]. In the following, we will discuss these approaches and highlight their differences particularly in terms of whether auxiliary sensor noise is considered, in terms of training data (and simulation thereof), and sensor setups different from the scenario considered in this thesis.

As discussed for single-channel approaches, several multi-channel approaches employ vibration sensors or accelerometers as the auxiliary sensor, and **it is assumed that the sensor does not pick up environmental noise** [17, 37, 38, 84, 85, 87]. This presents a limitation when considering in-ear microphones.

Similar to the single-channel approaches discussed in the previous section, several multi-channel approaches were only trained and evaluated on recorded signals of a single talker [37, 87], which may limit **generalization to different talkers**. In [37], fully convolutional DNN-based approaches were proposed that combine noisy outer microphone and auxiliary sensor signals by early fusion (in the first layer of the DNN) or late fusion (in the last layer). Experiments showed that multi-channel approaches using either of these fusion strategies, substantially outperformed comparable single-channel approaches using either only the outer microphone or only the auxiliary sensor signal. A limitation of the study in [37] is that environmental noise at the auxiliary sensor was not considered in the experiments. Additionally, both training and evaluation were carried out using recorded signals of the same single talker. This presents another limitation, since this approach may not generalize to different talkers. While personalization of OVR may be able to account of individual differences, it is unclear how this training-based personalization would

perform in comparison to generic (non-personalized) approaches trained for many talkers.

However, the ability to generalize to unseen talkers may require training with recorded signals from many talkers, as in [84]. In order to address **limited availability of training data**, other approaches simulate auxiliary sensor signals [8, 17, 38, 82, 86]. In [84], a multi-channel OVR approach was proposed for an outer microphone and a vibration sensor. An experimental evaluation was performed with recorded signals for training and testing, assuming no environmental noise at the vibration sensor. Even for unseen talkers, the proposed system improved speech quality compared to the baseline systems [27] and [37]. However, recording a sufficient amount of training data can be costly and time-consuming, which is why many other approaches rely on simulating auxiliary sensor signals. Building on the sensor fusion techniques in [37], an attention-based fusion technique for air- and body-conducted speech was proposed in [85]. This technique outperformed previous fusion techniques in experiments using recorded signals. Similarly as in [37] environmental noise was not considered at the auxiliary sensor. In [38], the approach from [85] was extended by adding GAN-based training. Although the GAN-based training reduced recorded data requirements while maintaining performance, the DNN simulating body-conducted speech at the auxiliary sensor was not evaluated in terms of simulation accuracy. In [17], a multi-channel OVR system was trained as a GAN that takes signals of an accelerometer and an outer microphone as input. In order to simulate training data, a DNN was trained to estimate accelerometer signals from outer microphone signals first. Experiments on simulated and recorded signals (assuming no environmental noise at the accelerometer) showed that using the accelerometer helps to suppress noise and interfering talkers. However, the DNN-based simulation method for simulating the accelerometer signals itself was not directly evaluated in terms of simulation accuracy. In [82], a system was proposed that uses both an outer microphone and an accelerometer as input for OVR. The system was trained with simulated accelerometer signals using a simulation method similar to [74], and further fine-tuned with user-specific recorded signals to improve performance. Although the number of DNN parameters was not reported, the system was shown to operate in real time on different devices. As previously discussed for [74], the simulation method assumes linear time-invariant relative transfer functions, which may not accurately reflect own voice transmission. In [8], an OVR system using an outer and an auxiliary sensor was proposed. The system computes a time-frequency mask for the outer microphone and uses a VAD based on the auxiliary sensor signal to decide whether to pass enhanced or unprocessed outer microphone speech. The system was trained on simulated data based on convolutive transfer functions and then fine-tuned with recorded data, achieving quality improvements compared to unprocessed signals and compared to a multi-channel baseline system under real-time constraints. However, it is unclear whether the convolutive transfer function model was able to accurately predict own voice signals at the auxiliary sensor. In [86], a real-time OVR system using both an outer and an in-ear microphone was pretrained with simulated data and fine-tuned with recorded data. Experiments were conducted using recorded own voice and environmental noise for testing, although in-ear environmental noise was ap-

proximated by applying a broadband gain to the outer microphone noise signal. While this training considers transmission of environmental noise to the in-ear microphone, a broadband gain does not realistically model individual differences in device fit. The system considerably improved speech quality compared to a baseline system using only an outer microphone. Different methods for simulating own voice or environmental noise at an auxiliary sensor have been proposed, but often considerable gains from additional fine-tuning are observed. As already discussed for simulation methods in the context of single-channel OVR approaches, considering factors that influence own voice at an in-ear microphone, such as individual differences or phonemes being uttered, could improve training data quality and OVR performance.

While OVR approaches for practical applications should aim to perform **causal processing** in order to avoid large processing delays, this is not always the case: In [41], an OVR system using both an outer and an in-ear microphone was proposed. The system contains magnitude and phase enhancement modules, which consist mostly of convolutional layers, but also bidirectional LSTM layers for modeling temporal patterns. The training data, more in particular the in-ear microphone signals, were simulated using a DNN trained to estimate noisy in-ear speech signals from noisy outer microphone speech signals. Additional generic and personalized fine-tuning using recorded own voice signals of the target talker was reported to improve the performance of the system pretrained with simulated data. While this approach also considers individual differences in in-ear own voice signals, the approach requires non-causal processing, which presents a limitation for real-time applications.

Also, approaches to OVR were proposed employing **sensor configurations different from the scenario considered in this thesis**. In [87], a multi-channel approach using two in-ear microphones was trained to reconstruct own voice from recorded signals of a single talker in a noiseless environment. The approach used a fully convolutional network architecture with a custom learned filterbank input layer. Experimental results with recorded signals of the same talker showed that using two in-ear microphones improved performance compared to using one in-ear microphone. This result may not generalize to recorded signals of different talkers. While this approach also uses multiple sensors, it is not applicable to the scenario considered in this thesis, where only one outer and one in-ear microphone are available. In addition, this approach does not employ an outer microphone recording broadband speech, so that it is unclear whether the learned filterbank would also be helpful in the scenario considered in this thesis. In [81], a low-complexity OVR system for a throat vibration sensor and a close-talk microphone was proposed. In a systematic comparison using recorded signals and taking into account environmental noise at the auxiliary sensor, this approach outperformed those in [37, 40, 84], and performed similarly to the system in [38] (after adjusting the systems in [38] and [84] to have comparable computational complexity). However, it is not clear how this approach would perform if an outer microphone and an in-ear microphone of a hearable were used instead. In [89], an OVR system was proposed that used a close-talk microphone and an accelerometer. Reconstruction is performed on the accelerometer signal based on harmonicity features and magnitude spec-

Table 1.2: Overview of multi-channel DNN-based OVR approaches. IM noise: In-ear microphone (environmental) noise, FT: Fine-tuning with recorded data.

1st Author	Year	Ref.	Training data	FT	Test data	IM noise	Parameters
Tagliasacchi	2020	[17]	simulated	-	sim. & rec.	-	8.3 M ¹
Yu	2020	[37]	recorded ²	-	recorded ²	-	1.03 M ¹
Liu	2020	[87]	recorded ²	-	recorded ²	-	-
Wang	2022	[84]	recorded	-	recorded	-	-
Wang	2022	[38, 85]	recorded	-	recorded	-	5.84 M
He	2023	[82]	simulated	✓	recorded	-	3.1 M ¹
Han	2024	[86]	simulated	✓	recorded	(✓) ³	-
Ma	2024	[41]	simulated	✓	recorded	✓	-
Kuang	2024	[81]	recorded	-	recorded	✓	1 M
Heitkaemper	2025	[8]	simulated	✓	recorded	-	-
Song	2025	[89]	recorded	-	recorded	-	93 K
Li	2025	[88]	recorded	-	recorded	-	-

¹ Parameter count according to [81].

² Recordings of a single talker were used.

³ IM environmental noise was approximated by scaling OM noise with a broadband gain.

togram estimation, and the close-talk microphone is only used for phase estimation. Trained and evaluated using recorded data, the system performed better than [82] while requiring far less computation. In [88], a UNet-based OVR architecture was proposed that uses separate encoders for a close-talk microphone and a vibration sensor placed on the cranial vertex (on top of the head). An experimental comparison with recorded data showed that this approach achieved higher objective scores than [37, 38, 79, 84], and received the highest subjective ratings compared to the other baselines in a listening test. It should be noted that close-talk microphones as considered in [81, 88, 89] are not available in the scenario considered in this thesis, and that auxiliary sensors at the throat or the top of the head are more intrusive to the wearer than those included in a hearable.

From this overview, it becomes clear that the multi-channel DNN-based approaches discussed in this section share limitations with the single-channel DNN-based approaches discussed in Section 1.2.2. Most prominently, many approaches do not consider environmental noise at the auxiliary sensor [17, 37, 38, 84, 85, 87]. However, hearables are frequently used in noisy environments, which present a limitation to applying these approaches to hearables with in-ear microphones. Moreover, many methods that do employ simulated data use simplified models of the auxiliary sensor signals, and therefore do not consider important factors like individual differences and time-varying transmission. Since several approaches were proposed and validated on different sensor setups, for example considering close-talk microphones or vibration sensors at the throat or the top of the head, they might not be suited

for the scenario considered in this thesis, where only an outer microphone and an in-ear microphone of a hearable device are available.

1.3 Thesis outline and main contributions

The main objective of this thesis is to develop and evaluate multi-channel DNN-based OVR approaches which jointly process the outer and in-ear microphone signals of a hearable in noisy environments. The main focus is to improve speech quality and reduce DNN training data requirements by using models of the own voice transfer characteristics between the outer and the in-ear microphone. A second focus is to investigate the possible benefit of personalizing OVR systems to the individual talker. Finally, we also aim to explore the trade-off between performance and computational complexity, and to identify suitable objective metrics for OVR approaches that correlate well with subjective quality ratings.

This thesis makes three main contributions. First, **we propose a phoneme-dependent own voice transfer characteristics model that accounts for changes in the transmission path during speech production**. Experimental results on recorded own voice signals from 18 talkers demonstrate that the proposed model predicts in-ear own voice signals more accurately than phoneme-independent models. In addition, talker-averaged models achieve lower prediction errors for unseen talkers than individual models. Second, **we propose data augmentation techniques to train an OVR system which jointly processes the outer and in-ear microphone signals**. More in particular, the phoneme-dependent model of own voice transfer characteristics is used to simulate a large amount of in-ear own voice signals from a dataset of clean speech signals. For a generic (non-personalized) OVR system, experimental results show that phoneme-dependent individual augmentation yields better performance than phoneme-independent and talker-averaged augmentation, and outperforms training directly with the limited amount of recorded own voice signals. Moreover, additional fine-tuning with recorded own voice signals after training with augmented own voice signals significantly improves performance, even when a limited amount of recorded own voice signals is available. In addition, we investigate low-complexity OVR system variants to meet computational requirements. Third, we investigate **personalization of OVR systems, either via training-based personalization or enrollment-based personalization**. Training-based personalization is performed using personalized data augmentation, personalized fine-tuning, or a combination of both. Results from a listening test show that generic (non-personalized) OVR systems substantially improve subjective quality compared to unprocessed noisy outer microphone signals, with personalization providing an additional benefit for some talkers. Although many objective metrics do not correlate well with the subjective quality ratings, the intrusive ESTOI, the intrusive GPSM^q, and the non-intrusive LEAP show a high correlation. In addition to training-based personalization, we also investigate enrollment-based personalization, where only a single enrollment utterance of the talker recorded with the in-ear microphone is available. Experimental results on the Vibravox dataset [43] demonstrate that enrollment-based personalization yields

large performance gains over generic OVR, especially in scenarios with competing talkers, and is robust to dataset mismatch.

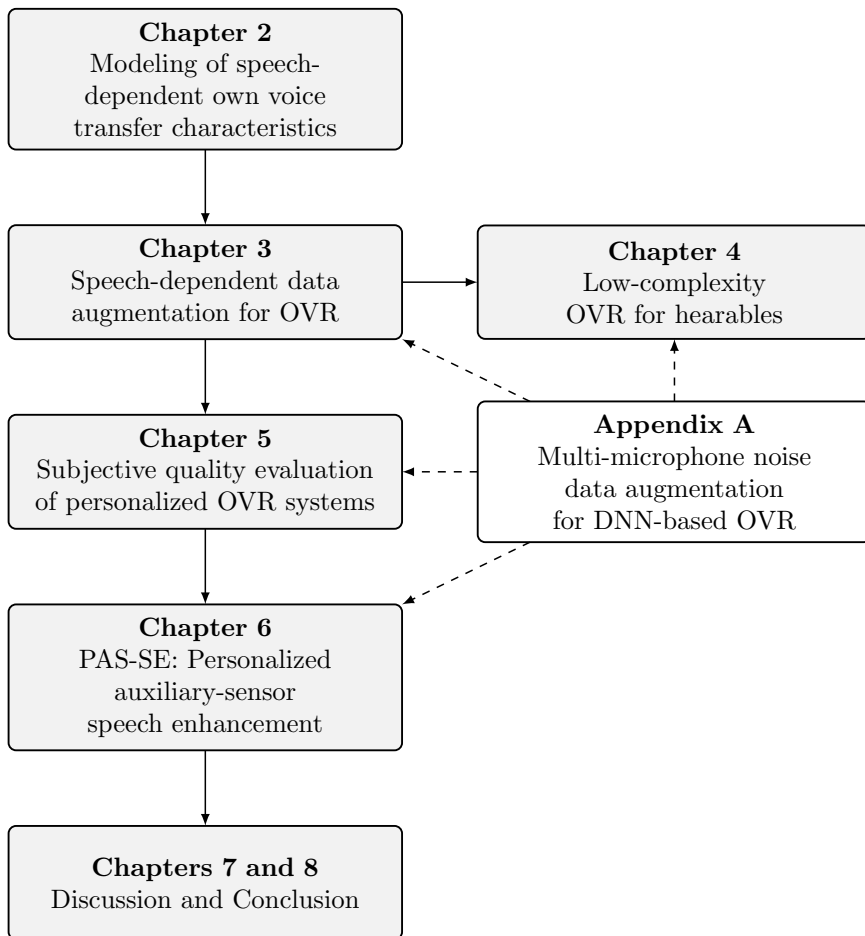


Figure 1.3: Thesis structure. The main chapters of the thesis are shown as gray boxes, while the appendix is shown as white box. The main order of chapters is illustrated by solid arrows, while dashed arrows illustrate prerequisites.

In the remainder of this section, we provide a chapter-by-chapter overview of this thesis. Figure 1.3 shows the thesis structure.

In **Chapter 2**, we propose a phoneme-dependent model of the own voice transfer characteristics between the outer and the in-ear microphone, assuming a linear time-invariant RTF for each phoneme. To estimate these RTFs, a small amount of recorded own voice signals are separated into different phoneme segments based on a phoneme recognition system. From these segments, a separate RTF for each phoneme is estimated between the outer and the in-ear microphone. To simulate

an in-ear microphone own voice signal from an outer microphone own voice signal, phoneme annotation is applied to the outer microphone signal and used to select the corresponding RTF for each segment. The in-ear own voice signal is simulated by applying (time-varying) phoneme-dependent filtering to the outer microphone signal. Experimental results on recorded own voice signals from 18 talkers with 306 utterances per talker (which we published on Zenodo) demonstrate that the proposed phoneme-dependent model is able to predict in-ear own voice signals more accurately than phoneme-independent models [39, 74, 78, 82]. In terms of technical metrics, the prediction error was reduced by up to 50% compared to phoneme-independent models. We consider both individual models and talker-averaged models. While individual models achieve lower prediction errors for matched talkers than talker-averaged models, talker-averaged models achieve lower prediction errors for unseen (mismatched) talkers. Chapter 2 was published in [90].

Aiming at reducing training data requirements for DNN-based OVR systems, in **Chapter 3** we propose to perform data augmentation using the phoneme-dependent own voice transfer characteristic models from Chapter 2. In this chapter, we consider an OVR system based on the frequency-and-time joint non-linear filter (FT-JNF) architecture [64], which computes complex-valued masks for the outer and in-ear microphone signals. In addition, we only consider generic OVR systems, designed to reconstruct own voice of many different talkers, instead of personalized systems. Whereas recording many utterances from different talkers wearing a specific hearable device is costly and time-consuming, the proposed augmentation only requires a limited amount of recorded own voice signals. The recorded own voice signals are used to compute an own voice transfer characteristics model, using which a large amount of in-ear own voice signals can be simulated from a clean speech dataset. In particular, we compare phoneme-dependent and phoneme-independent augmentation, using either individual or talker-averaged transfer characteristics models. For the experimental evaluation, we consider the same own voice dataset as in the previous chapter, consisting of recordings from 18 talkers with 306 utterances per talker. The OVR performance is evaluated for diverse environmental noise types, at SNRs ranging between -10 dB and 10 dB at the outer microphone. To obtain multi-microphone environmental noise signals for training and evaluation, we use a spatialization procedure that uses individually measured transfer functions from eight loudspeakers (more details in Appendix A). Results show that training with phoneme-dependent individual augmentation followed by additional fine-tuning with recorded signals yields the best performance, achieving improvements of about 1.3 in terms of PESQ and over 0.15 in terms of STOI compared to the noisy outer microphone signal. We also investigate the trade-off between OVR performance and the amount of recorded own voice signals, both in terms of number of talkers and number of utterances per talker. Results show that using recorded data from 8 talkers with 306 utterances or 12 talkers with 12 utterances each yield similar performance as using all recorded data for augmentation and fine-tuning. Chapter 3 was published in [91].

In **Chapter 4**, we investigate low-complexity variants of the OVR system from Chapter 3 trained using phoneme-dependent data augmentation. We consider vari-

ants with different computational complexity, ranging from the system in the previous chapter (1.39 M parameters, 22.38 GMACs/s), down to a system with heavily reduced complexity (13 k parameters, 0.23 GMACs/s). Results show that the OVR performance decreases when model size decreases, where the smallest considered variant still achieves a PESQ improvement of 0.7 compared to the noisy outer microphone signal. In addition, the results demonstrate that the considered variants outperform baseline OVR systems [74, 75, 91, 92] at a comparable complexity, also when using a small amount of recorded data for augmentation and fine-tuning. Chapter 4 was published in [93].

In **Chapter 5**, we propose to personalize the OVR system from Chapter 3 via training-based personalization, i.e., either using personalized data augmentation, personalized fine-tuning, or a combination of both. Instead of using transfer characteristics models and fine-tuning utterances of many talkers, training-based personalization uses transfer characteristic models and fine-tuning utterances of a single talker to obtain a personalized system. We evaluate the benefit of generic and personalized OVR systems through a listening test and compare objective metric predictions with subjective quality ratings. Results from a listening test show that generic (non-personalized) OVR systems substantially improve subjective quality compared to unprocessed noisy outer microphone signals with an average score improvement of 50 MUSHRA points, with personalization providing an additional benefit of up to 5 points for some talkers. A comparison of subjective quality ratings with predictions by objective metrics provides further insights into which metrics are suitable to predict OVR performance, with particularly good correlations being observed for the metrics ESTOI ($r = 0.89$, $\rho_S = 0.92$), GPSM^q ($r = -0.86$, $\rho_S = -0.88$), and LEAP ($r = -0.86$, $\rho_S = -0.87$). Chapter 5 is based on [94].

Whereas Chapter 5 proposed training-based personalization of OVR systems, in **Chapter 6** we explore enrollment-based personalization, requiring only a single enrollment utterance of the talker. For enrollment-based personalization, the OVR system is trained with own voice signals and enrollment signals of many different talkers, aiming at generalizing to unseen talkers. Different from previous chapters, the in-ear microphone is not just considered as an input for processing, but also for recording the enrollment utterance. Experimental results on the Vibravox [43] dataset demonstrate that enrollment-based personalization is very effective for OVR in the presence of interfering talkers, achieving up to 10 dB SI-SDR improvement over unprocessed signals. Results also show that the proposed enrollment-based personalization consistently improves OVR performance even under dataset mismatch (testing on the recorded dataset considered in previous chapters) and with noisy enrollment utterances. Chapter 6 is based on [95].

In **Chapter 7**, we discuss the main contributions from Chapters 2 to 6 and outline possible directions for further research.

References

- [1] C.-Y. Chang, A. Siswanto, C.-Y. Ho, T.-K. Yeh, Y.-R. Chen, and S. M. Kuo, “Listening in a noisy environment: Integration of active noise control in audio

- products,” *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 34–43, 2016. DOI: [10.1109/MCE.2016.2590159](#).
- [2] H. Schepker, F. Denk, B. Kollmeier, and S. Doclo, “Acoustic transparency in hearables - perceptual sound quality evaluations,” *Journal of the Audio Engineering Society*, vol. 68, no. 7/8, pp. 495–507, 2020. DOI: [10.17743/jaes.2020.0045](#).
 - [3] V. Välimäki, A. Franck, J. Rämö, H. Gamper, and L. Savioja, “Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 92–99, 2015. DOI: [10.1109/MSP.2014.2369191](#).
 - [4] R. Gupta, J. He, R. Ranjan, W.-S. Gan, F. Klein, C. Schneiderwind, A. Neidhardt, K. Brandenburg, and V. Välimäki, “Augmented/mixed reality audio for hearables: Sensing, control, and rendering,” *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 63–89, 2022. DOI: [10.1109/MSP.2021.3110108](#).
 - [5] R. E. Bouserhal, T. H. Falk, and J. Voix, “Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones,” in *Proc. Meetings on Acoustics (ICA)*, vol. 19, Montreal, QC, Canada, 2013. DOI: [10.1121/1.4800452](#).
 - [6] S. Nordholm, A. Davis, P. C. Yong, and H. H. Dam, “Assistive listening headsets for high noise environments: Protection and communication,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Apr. 2015, pp. 5753–5757. DOI: [10.1109/ICASSP.2015.7179074](#).
 - [7] I. López-Espejo, E. Roselló, A. Edraki, N. Harte, and J. Jensen, “Noise-robust hearing aid voice control,” *IEEE Signal Process. Lett.*, vol. 32, pp. 241–245, 2025. DOI: [10.1109/LSP.2024.3512377](#).
 - [8] J. Heitkaemper, J. Caroselli, M. McKinnon, A. Narayanan, and N. Howard, “Bone conducted signal guided speech enhancement for voice assistant on earbuds,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025. DOI: [10.1109/ICASSP49660.2025.10889416](#).
 - [9] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015. DOI: [10.1109/MSP.2014.2366780](#).
 - [10] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017. DOI: [10.1109/TASLP.2017.2726762](#).
 - [11] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A consolidated perspective on multimicrophone speech enhancement and source separation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017. DOI: [10.1109/TASLP.2016.2647702](#).

- [12] M. Stamenovic, N. L. Westhausen, L.-C. Yang, C. Jensen, and A. Pawlicki, “Weight, Block or Unit? Exploring Sparsity Tradeoffs for Speech Enhancement on Tiny Neural Accelerators,” in *Proc. NeurIPS Workshop Efficient Natural Language and Speech Processing*, Nov. 2021. DOI: [10.48550/arXiv.2111.02351](#).
- [13] P. Hoang, J. M. de Haan, Z.-H. Tan, and J. Jensen, “Multichannel speech enhancement with own voice-based interfering speech suppression for hearing assistive devices,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 706–720, 2022. DOI: [10.1109/TASLP.2022.3145294](#).
- [14] H. Schröter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, “DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7407–7411. DOI: [10.1109/ICASSP43922.2022.9747055](#).
- [15] H. Bae, P. Andreev, A. Saginbaev, N. Babaev, W. Lee, H. Sung, and H.-Y. Cho, “Speech boosting: Low-latency live speech enhancement for TWS earbuds,” in *Proc. Interspeech*, Kos, Greece, Sep. 2024, pp. 647–651. DOI: [10.21437/Interspeech.2024-1444](#).
- [16] T. Röddiger, C. Clarke, P. Breitling, T. Schneegans, H. Zhao, H. Gellersen, and M. Beigl, “Sensing with earables: A systematic literature review and taxonomy of phenomena,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 3, Sep. 2022. DOI: [10.1145/3550314](#).
- [17] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, “SEANet: A multi-modal speech enhancement network,” in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1126–1130. DOI: [10.21437/Interspeech.2020-1563](#).
- [18] Y. Sui, M. Zhao, J. Xia, X. Jiang, and 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,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 4, Nov. 2024. DOI: [10.1145/3699757](#).
- [19] F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J.-H. Bach, J. Wellmann, and B. Kollmeier, “A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research,” in *Proc. AES International Conference on Headphone Technology*, San Francisco, USA, Aug. 2019.
- [20] T. Röddiger, J. Stuchbury-Wass, M. Ciliberto, P. Lepold, and M. Beigl, “OpenEarable 1.4: Dual microphones earpiece to capture in-ear and outer-ear audio signals,” in *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Melbourne, Australia, 2024, pp. 930–933. DOI: [10.1145/3675094.3678483](#).
- [21] D. Shi, B. Lam, K. Ooi, X. Shen, and W.-S. Gan, “Selective fixed-filter active noise control based on convolutional neural network,” *Signal Processing*, vol. 190, p. 108317, 2022. DOI: [10.1016/j.sigpro.2021.108317](#).

- [22] T. Xiao and S. Doclo, “Effect of target signals and delays on spatially selective active noise control for open-fitting hearables,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 1056–1060. DOI: [10.1109/ICASSP48485.2024.10445843](https://doi.org/10.1109/ICASSP48485.2024.10445843).
- [23] S. Liebich and P. Vary, “Occlusion Effect Cancellation in Headphones and Hearing Devices—The Sister of Active Noise Cancellation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 35–48, 2022. DOI: [10.1109/TASLP.2021.3130966](https://doi.org/10.1109/TASLP.2021.3130966).
- [24] T. Sankowsky-Rothe, M. Blau, S. Köhler, and A. Stirnemann, “Individual equalization of hearing aids with integrated ear canal microphones,” *Acta Acustica united with Acustica*, vol. 101, no. 3, pp. 552–566, 2015. DOI: [10.3813/AAA.918852](https://doi.org/10.3813/AAA.918852).
- [25] S. Vogl and M. Blau, “Individualized prediction of the sound pressure at the eardrum for an earpiece with integrated receivers and microphones,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 917–930, Feb. 2019. DOI: [10.1121/1.5089219](https://doi.org/10.1121/1.5089219).
- [26] R. E. Bouserhal, A. Bernier, and J. Voix, “An in-ear speech database in varying conditions of the audio-phonation loop,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 1069–1077, Feb. 2019. DOI: [10.1121/1.5091777](https://doi.org/10.1121/1.5091777).
- [27] K. Kondo, T. Fujita, and K. Nakagawa, “On equalization of bone conducted speech for improved speech quality,” in *Proc. International Symposium on Signal Processing and Information Technology*, Vancouver, BC, Canada, Aug. 2006, pp. 426–431. DOI: [10.1109/ISSPIT.2006.270839](https://doi.org/10.1109/ISSPIT.2006.270839).
- [28] S. Stenfelt and S. Reinfeldt, “A model of the occlusion effect with bone-conducted stimulation,” *International Journal of Audiology*, vol. 46, no. 10, pp. 595–608, Jan. 2007. DOI: [10.1080/14992020701545880](https://doi.org/10.1080/14992020701545880).
- [29] M. Blau, R. Roden, N. Hauenschild, S. Kersten, R. Rehman, M. Vorländer, and J. Fels, “Methods to experimentally characterize the own-voice-generated objective occlusion effect induced by hearables,” *Acta Acustica*, vol. 9, p. 73, 2025. DOI: [10.1051/aacus/2025055](https://doi.org/10.1051/aacus/2025055).
- [30] S. Reinfeldt, P. Östli, B. Håkansson, and S. Stenfelt, “Hearing one’s own voice during phoneme vocalization - Transmission by air and bone conduction,” *J. Acoust. Soc. Am.*, vol. 128, no. 2, pp. 751–762, Aug. 2010. DOI: [10.1121/1.3458855](https://doi.org/10.1121/1.3458855).
- [31] H. Saint-Gaudens, H. Nélisse, F. Sgard, and O. Doutres, “Towards a practical methodology for assessment of the objective occlusion effect induced by earplugs,” *J. Acoust. Soc. Am.*, vol. 151, no. 6, pp. 4086–4100, Jun. 2022. DOI: [10.1121/10.0011696](https://doi.org/10.1121/10.0011696).
- [32] J. Richard, V. Zimpfer, C. Blondé-Weinmann, and S. Roth, “Change in transfer function between air and bone conduction microphones due to mouth opening variation,” *Applied Acoustics*, vol. 228, p. 110 293, Jan. 2025. DOI: <https://doi.org/10.1016/j.apacoust.2024.110293>.
- [33] M. Ø. Hansen, “Occlusion effects Part I and II,” PhD thesis, Department of Acoustic Technology, Technical University of Denmark, 1998.

- [34] M. K. Brummund, F. Sgard, Y. Petit, and F. Laville, “Three-dimensional finite element modeling of the human external ear: Simulation study of the bone conduction occlusion effect,” *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1433–1444, Mar. 2014. DOI: [10.1121/1.4864484](#).
- [35] T. Dekens and W. Verhelst, “Body conducted speech enhancement by equalization and signal fusion,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2481–2492, 2013. DOI: [10.1109/TASL.2013.2274696](#).
- [36] R. E. Bouserhal, T. H. Falk, and J. Voix, “In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension,” *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 1321–1331, Mar. 2017. DOI: [10.1121/1.4976051](#).
- [37] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, “Time-domain multi-modal bone/air conducted speech enhancement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020. DOI: [10.1109/LSP.2020.3000968](#).
- [38] H. Wang, X. Zhang, and D. Wang, “Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 3134–3143, 2022. DOI: [10.1109/TASLP.2022.3209943](#).
- [39] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, “Configurable EBEN: Extreme bandwidth extension network to enhance body-conducted speech capture,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 3499–3512, 2023. DOI: [10.1109/TASLP.2023.3313433](#).
- [40] C. Zheng, L. Xu, X. Fan, J. Yang, J. Fan, and X. Huang, “Dual-path transformer-based network with equalization-generation components prediction for flexible vibrational sensor speech enhancement in the time domain,” *J. Acoust. Soc. Am.*, vol. 151, no. 5, pp. 2814–2825, Apr. 2022. DOI: [10.1121/10.0010316](#).
- [41] D. Ma, T. Dang, M. Ding, and R. Balan, “Clearspeech: Improving voice quality of earbuds using both in-ear and out-ear microphones,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 4, Jan. 2024. DOI: [10.1145/3631409](#).
- [42] B. Gårdbæk and P. Kidmose, “On the origin of cardiovascular sounds recorded from the ear,” *IEEE Trans. Biomed. Eng.*, vol. 72, no. 1, pp. 210–216, 2025. DOI: [10.1109/TBME.2024.3445412](#).
- [43] J. Hauret, M. Olivier, T. Joubaud, C. Langrenne, S. Poirée, V. Zimpfer, and É. Bavu, “Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors,” *Speech Communication*, Apr. 2025. DOI: [10.1016/j.specom.2025.103238](#).
- [44] H. S. Shin, H.-G. Kang, and T. Fingscheidt, “Survey of speech enhancement supported by a bone conduction microphone,” in *Proc. ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012.

- [45] M. S. Rahman and T. Shimamura, “Intelligibility enhancement of bone conducted speech by an analysis-synthesis method,” in *Proc. International Midwest Symposium on Circuits and Systems (MWSCAS)*, Seoul, South Korea, Aug. 2011. DOI: [10.1109/MWSCAS.2011.6026374](#).
- [46] H. S. Shin, T. Fingscheidt, and H.-G. Kang, “A priori SNR estimation using air- and bone-conduction microphones,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2015–2025, Nov. 2015. DOI: [10.1109/TASLP.2015.2446202](#).
- [47] C.-H. Lee, B. D. Rao, and H. Garudadri, “Bone-conduction sensor assisted noise estimation for improved speech enhancement,” in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 1180–1184. DOI: [10.21437/Interspeech.2018-1046](#).
- [48] M. Tammen, X. Li, S. Doclo, and L. Theverapperuma, “Dictionary-Based Fusion of Contact and Acoustic Microphones for Wind Noise Reduction,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022. DOI: [10.1109/IWAENC53105.2022.9914710](#).
- [49] A. Lu, K. Sarkar, M. Mittal, R. M. Corey, P. Smaragdis, and A. C. Singer, “Denoising bandwidth extension via fusion of air and bone microphones,” in *Proc. Meetings on Acoustics*, vol. 51, Chicago, Illinois, USA, May 2023. DOI: [10.1121/2.0002023](#).
- [50] R. M. Corey, “Mixed-delay distributed beamforming for own-speech separation in hearing devices with wireless remote microphones,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2023. DOI: [10.1109/WASPAA58266.2023.10248172](#).
- [51] W. Middelberg, J.-S. Lee, S. B. Sereshki, A. Aroudi, V. Tourbabin, and D. D. E. Wong, “Microphone occlusion mitigation for own-voice enhancement in head-worn microphone arrays using switching-adaptive beamforming,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Tahoe City, CA, USA, Oct. 2025. DOI: [10.1109/WASPAA66052.2025.11230992](#).
- [52] J. Abel and T. Fingscheidt, “Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2018. DOI: [10.1109/TASLP.2017.2761236](#).
- [53] H. Wang and D. Wang, “Time-frequency loss for CNN based speech super-resolution,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 861–865. DOI: [10.1109/ICASSP40776.2020.9053712](#).
- [54] J. Su, Y. Wang, A. Finkelstein, and Z. Jin, “Bandwidth extension is all you need,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 696–700. DOI: [10.1109/ICASSP39728.2021.9413575](#).

- [55] H. Wang and D. Wang, “Towards robust speech super-resolution,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 2058–2066, 2021. DOI: [10.1109/TASLP.2021.3054302](#).
- [56] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, “HIFI++: A unified framework for bandwidth extension and speech enhancement,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 2023. DOI: [10.1109/ICASSP49357.2023.10097255](#).
- [57] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 31–35. DOI: [10.1109/ICASSP.2016.7471631](#).
- [58] Z.-Q. Wang, P. Wang, and D. Wang, “Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 1778–1787, 2020. DOI: [10.1109/TASLP.2020.2998279](#).
- [59] W. Mack and E. A. P. Habets, “Deep filtering: Signal extraction and reconstruction using complex time-frequency filters,” *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2020. DOI: [10.1109/LSP.2019.2955818](#).
- [60] N. L. Westhausen and B. T. Meyer, “Dual-signal transformation LSTM network for real-time noise suppression,” in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 2477–2481. DOI: [10.21437/Interspeech.2020-2631](#).
- [61] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, “Towards efficient models for real-time deep noise suppression,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 656–660. DOI: [10.1109/ICASSP39728.2021.9413580](#).
- [62] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, “TF-GridNet: Integrating full- and sub-band modeling for speech separation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023. DOI: [10.1109/TASLP.2023.3304482](#).
- [63] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, “STFT-domain neural speech enhancement with very low algorithmic latency,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 397–410, 2023. DOI: [10.1109/TASLP.2022.3224285](#).
- [64] K. Tesch and T. Gerkmann, “Insights into deep non-linear filters for improved multi-channel speech enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023. DOI: [10.1109/TASLP.2022.3221046](#).
- [65] R. Sinha, C. Rollwage, and S. Doclo, “Low-complexity real-time single-channel speech enhancement based on skip-GRUs,” in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, Sep. 2023, pp. 181–185. DOI: [10.30420/456164035](#).

- [66] M. Tammen and S. Doclo, “Parameter estimation procedures for deep multi-frame MVDR filtering for single-microphone speech enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 3237–3248, 2023. DOI: [10.1109/TASLP.2023.3306715](#).
- [67] R. Haeb-Umbach, T. Nakatani, M. Delcroix, C. Boeddeker, and T. Ochiai, “Microphone array signal processing and deep learning for speech enhancement: Combining model-based and data-driven approaches to parameter estimation and filtering,” *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 12–23, 2024. DOI: [10.1109/MSP.2024.3451653](#).
- [68] N. L. Kühne, J. Østergaard, J. Jensen, and Z.-H. Tan, “xLSTM-SENet: xLSTM for single-channel speech enhancement,” in *Proc. Interspeech*, Rotterdam, The Netherlands, Aug. 2025, pp. 5148–5152. DOI: [10.21437/Interspeech.2025-108](#).
- [69] S. Araki, N. Ito, R. Haeb-Umbach, G. Wichern, Z.-Q. Wang, and Y. Mitsufuji, “30+ years of source separation research: Achievements and future challenges,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025. DOI: [10.1109/ICASSP49660.2025.10889006](#).
- [70] S. S. Shetu, E. A. P. Habets, and A. Brendel, “Comparative analysis of discriminative deep learning-based noise reduction methods in low snr scenarios,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aalborg, Denmark, Sep. 2024, pp. 36–40. DOI: [10.1109/IWAENC61483.2024.10694283](#).
- [71] H.-P. Liu, Y. Tsao, and C.-S. Fuh, “Bone-conducted speech enhancement using deep denoising autoencoder,” *Speech Communication*, vol. 104, pp. 106–112, Nov. 2018. DOI: [10.1016/j.specom.2018.06.002](#).
- [72] H. Park, Y.-S. Shin, and S.-H. Shin, “Speech quality enhancement for in-ear microphone based on neural network,” *IEICE Trans. on Information and Systems*, vol. 102, no. 8, pp. 1594–1597, 2019. DOI: [10.1587/transinf.2018EDL8249](#).
- [73] H. Q. Nguyen and M. Unoki, “Improvement in bone-conducted speech restoration using linear prediction and long short-term memory model,” *Journal of Signal Processing*, vol. 24, no. 4, pp. 175–178, 2020. DOI: [10.2299/jsp.24.175](#).
- [74] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Training strategies for own voice reconstruction in hearing protection devices using an in-ear microphone,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022. DOI: [10.1109/IWAENC53105.2022.9914801](#).
- [75] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, “EBEN: Extreme bandwidth extension network applied to speech signals captured with noise-resilient body-conduction microphones,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, Jun. 2023. DOI: [10.1109/ICASSP49357.2023.10096301](#).

- [76] Y. Li, Y. Wang, X. Liu, Y. Shi, S. Patel, and S.-F. Shih, “Enabling real-time on-chip audio super resolution for bone-conduction microphones,” *Sensors*, vol. 23, no. 1, Jan. 2023. DOI: [10.3390/s23010035](#).
- [77] C. Li, F. Yang, and J. Yang, “A two-stage approach to quality restoration of bone-conducted speech,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 818–829, 2024. DOI: [10.1109/TASLP.2023.3337988](#).
- [78] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, “Speaker adaptation for enhancement of bone-conducted speech,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 10 456–10 460. DOI: [10.1109/ICASSP48485.2024.10447322](#).
- [79] C. Li, F. Yang, and J. Yang, “Restoration of bone-conducted speech with U-Net-like model and energy distance loss,” *IEEE Signal Process. Lett.*, vol. 31, pp. 166–170, 2024. DOI: [10.1109/LSP.2023.3347149](#).
- [80] J. Hauret, T. Joubaud, and É. Bavu, “Real-time speech enhancement in noise for throat microphone using neural audio codec as foundation model,” *arXiv:2508.02974*, Aug. 2025. DOI: [10.48550/arXiv.2508.02974](#).
- [81] K. Kuang, F. Yang, and J. Yang, “A lightweight speech enhancement network fusing bone- and air-conducted speech,” *J. Acoust. Soc. Am.*, vol. 156, no. 2, pp. 1355–1366, Aug. 2024. DOI: [10.1121/10.0028339](#).
- [82] L. He, H. Hou, S. Shi, X. Shuai, and Z. Yan, “Towards bone-conducted vibration speech enhancement on head-mounted wearables,” in *Proc. Annual International Conference on Mobile Systems, Applications and Services*, New York, USA, Jun. 2023, pp. 14–27. DOI: [10.1145/3581791.3596832](#).
- [83] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, Oct. 2015, pp. 234–241. DOI: [10.1007/978-3-319-24574-4_28](#).
- [84] M. Wang, J. Chen, X. Zhang, Z. Huang, and S. Rahardja, “Multi-modal speech enhancement with bone-conducted speech in time domain,” *Applied Acoustics*, vol. 200, Nov. 2022. DOI: [10.1016/j.apacoust.2022.109058](#).
- [85] H. Wang, X. Zhang, and D. Wang, “Attention-based fusion for bone-conducted and air-conducted speech enhancement in the complex domain,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7757–7761. DOI: [10.1109/ICASSP43922.2022.9746374](#).
- [86] F. Han, P. Yang, Y. Zuo, F. Shang, F. Xu, and X.-Y. Li, “Earspeech: Exploring in-ear occlusion effect on earphones for data-efficient airborne speech enhancement,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 3, Sep. 2024. DOI: [10.1145/3678594](#).
- [87] C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, “Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1888–1900, 2020. DOI: [10.1109/TASLP.2020.2976193](#).

- [88] C. Li, F. Yang, and J. Yang, “Bone conduction-aided speech enhancement with two-tower network and contrastive learning,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 33, pp. 163–174, 2025. DOI: [10.1109/TASLP.2024.3512207](#).
- [89] Y. Song, Y. Kim, and Y. Chung, “Lightweight speech enhancement model based on harmonic attention and phase estimation with skin-attachable accelerometer,” in *Proc. Interspeech*, Rotterdam, The Netherlands, Aug. 2025, pp. 66–70. DOI: [10.21437/Interspeech.2025-2642](#).
- [90] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Modeling of speech-dependent own voice transfer characteristics for hearables with in-ear microphones,” *Acta Acustica*, vol. 8, p. 18, 2024. DOI: [10.1051/aacus/2024032](#).
- [91] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Speech-dependent data augmentation for own voice reconstruction with hearable microphones in noisy environments,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 8, no. 32, 2025. DOI: [10.1186/s13636-025-00418-1](#).
- [92] 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, Dec. 2023. DOI: [10.1109/OJSP.2023.3343320](#).
- [93] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Low-complexity own voice reconstruction for hearables with an in-ear microphone,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025. DOI: [10.1109/ICASSP49660.2025.10887874](#).
- [94] M. Ohlenbusch, C. Rollwage, S. Doclo, and J. Rannies, “Subjective quality evaluation of personalized own voice reconstruction systems,” *Acta Acustica*, vol. 10, no. 26, 2026. DOI: [10.1051/aacus/2026021](#).
- [95] M. Ohlenbusch, M. Kegler, and M. Stamenovic, “PAS-SE: Personalized auxiliary-sensor speech enhancement for voice pickup in hearables,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2026, pp. 18 942–18 946. DOI: [10.1109/ICASSP55912.2026.11460554](#).

MODELING OF SPEECH-DEPENDENT OWN VOICE TRANSFER CHARACTERISTICS FOR HEARABLES WITH AN IN-EAR MICROPHONE

This chapter is identical in content to the publication: M. Ohlenbusch, C. Rollwage, and S. Doclo, “Modeling of speech-dependent own voice transfer characteristics for hearables with an in-ear microphone,” *Acta Acustica*, vol. 8, no. 28, 2024. DOI: 10.1051/aacus/2024032. [Online]. Available: 10.1051/aacus/2024032.

Authors	Author's contribution							
	A	B	C	D	E	F	G	H
Mattes Ohlenbusch	✗	✗	✗	✗	✗	✗	✗	✗
Christian Rollwage	✗			✗		✗	✗	✗
Simon Doclo	✗			✗		✗	✗	✗

A - Substantial contributions to the conception or design of the work

B - Acquisition of the data

C - Analysis of the data

D - Interpretation of the data

E - Drafting the work

F - Revising the work critically

G - Final approval of the version to be published

H - Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Abstract

Many hearables contain an in-ear microphone, which may be used to capture the own voice of its user. However, due to the hearable occluding the ear canal, the in-ear microphone mostly records body-conducted speech, typically suffering from band-limitation effects and amplification at low frequencies. Since the occlusion effect is determined by the ratio between the air-conducted and body-conducted components of own voice, the own voice transfer characteristics between the outer face of the hearable and the in-ear microphone depend on the speech content and the individual talker. In this paper, we propose a speech-dependent model of the own voice transfer characteristics based on phoneme recognition, assuming a linear time-invariant relative transfer function for each phoneme. We consider both individual models and models averaged over several talkers. Experimental results based on recordings with a prototype hearable show that the proposed speech-dependent model enables to simulate in-ear signals more accurately than a speech-independent model in terms of technical measures, especially under utterance mismatch and talker mismatch. Additionally, simulation results show that talker-averaged models generalize better to different talkers than individual models.

2.1 Introduction

Hearables, i.e. smart earpieces containing a loudspeaker and one or more microphones, are often used for speech communication in noisy acoustic environments. In this paper, we consider the scenario where the hearable is used to pick up the own voice of the user talking in a noisy environment (e.g., to be transmitted via a wireless link to a mobile phone or another hearable). Assuming that the hearable is at least partly occluding the ear canal, in this scenario an in-ear microphone may be beneficial to pick up the own voice since environmental noise is attenuated. Compared to own voice recorded at the outer face of the hearable, own voice recorded inside an occluded ear is known to suffer from amplification at low frequencies (below ca. 1 kHz) and strong attenuation at higher frequencies (above ca. 2 kHz), leading to a limited bandwidth [1]. The occlusion effect is determined by the ratio between the air-conducted and body-conducted components of own voice, which depends on device properties such as earmould fit and insertion depth [2], individual anatomic factors such as residual ear canal volume and shape [3, 4], and the generated sounds or phonemes [5, 6]. In particular, it has been shown that the occlusion effect for different vowels can be predicted by a linear combination of their formant frequencies [7], with closed front vowels exhibiting the largest occlusion effect. In addition, mouth movements during articulation [8] and body-conduction from different places of excitation [9] likely influence the occlusion effect as well. Unlike acoustical models based on ear canal geometry [3] or three-dimensional finite element models of body-conduction occlusion [10], in this paper we consider a signal processing-based approach to model the own voice transfer characteristics between a microphone at the entrance of the occluded ear canal (i.e. at the outer face of the hearable) and an in-ear microphone.

In many hearable applications, acoustic transfer path models for the microphone inside the occluded ear canal are required. For example, active noise cancellation algorithms may benefit from an accurate estimate of the so-called secondary path between the hearable loudspeaker and the in-ear microphone [11, 12]. In active occlusion cancellation (AOC), models of the own voice transfer path between the microphones inside and outside the occluded ear canal can be used to generate a cancellation signal that aims at compensating the occlusion effect as measured at the in-ear microphone [13, 14]. Models of the own voice transfer path are not only relevant for AOC, but also for algorithms to enhance the quality of the in-ear microphone signal picking up the own voice of the user. Several own voice reconstruction algorithms aiming at bandwidth extension, equalization and noise reduction have been proposed, e.g., based on classical signal processing [15] or supervised learning [16–19]. Supervised learning-based approaches typically require large amounts of training data. Since large amounts of realistic in-ear recordings may be hard to obtain for several talkers, an accurate and possibly individual model of the own voice transfer characteristics would be highly beneficial. Such a model would enable to generate large amounts of simulated in-ear signals either from recordings at the entrance of the ear canal or from speech corpora, e.g., [20]. Data augmentation can then be performed with these simulated in-ear signals to train supervised learning-based own voice reconstruction algorithms. Similarly as

for other acoustic signal processing applications [21–23], it is expected that using more accurate acoustic models for generating augmented training data improves system performance and generalization ability.

Several models of own voice transfer characteristics have been presented in the literature, either between two air-conduction microphones [17] or between an air-conduction and a body-conduction microphone [16, 18, 24]. In [24], it has been proposed to convert air-conducted to bone-conducted speech using a deep neural network (DNN) model that accounts for individual differences between talkers based on a speaker identification system. In [16], a DNN model estimating bone-conducted speech from air-conducted speech is jointly trained with a multi-modal enhancement network within a semi-supervised training scheme, resulting in reduced data requirements compared to fully supervised training. Instead of using rather complicated black-box DNN models, in [17, 18] time-invariant linear relative transfer functions (RTFs) are used to model own voice transfer characteristics. To introduce variations in the simulated own voice signals, either RTFs estimated on recordings of multiple talkers are used [17], or random values are added to the magnitude of the RTF estimated from a single talker [18]. It should be realized that these variations do not account for the speech-dependent nature of the own voice transfer characteristics.

Aiming at obtaining a model of the own voice transfer characteristics that generalizes well to unseen utterances and talkers, in this paper we propose a speech-dependent system identification approach, where for each phoneme a different RTF between the microphone at the entrance of the occluded ear canal and the in-ear microphone is estimated. We consider both individual and talker-averaged models. To simulate in-ear own voice signals from broadband speech, a phoneme recognition system is first utilized to segment the broadband speech into different segments corresponding to a specific phoneme, which are then filtered using the corresponding (smoothed) phoneme-specific RTFs. In contrast to previous RTF-based modeling approaches [17, 18], the proposed model of own voice transfer characteristics is speech-dependent and thus time-varying. In addition, contrary to the DNN-based modeling approach [16], only a small amount of own voice recordings are required for model estimation. The accuracy of simulating in-ear signals is assessed using recorded own voice signals of over 300 utterances by 18 talkers, each wearing a prototype hearable device [25]. The role of speech-dependency for simulating in-ear own voice signals is investigated by comparing the proposed speech-dependent RTF-based model to a speech-independent RTF-based model, and an adaptive filtering-based model [26] which is utterance-specific. Experimental results show that the proposed speech-dependent model enables to simulate in-ear own voice signals more accurately than the speech-independent model and the adaptive filtering-based model in terms of technical distance measures. In addition, the performance of individual and talker-averaged models is compared in terms of their generalization capability to unseen talkers. Results show that the speech-dependent talker-averaged model generalizes better to utterances of unseen talkers compared to speech-independent or individual models. Preliminary results of the proposed approach have already been published in [27]. This paper extends upon previous work presented in [27] by proposing

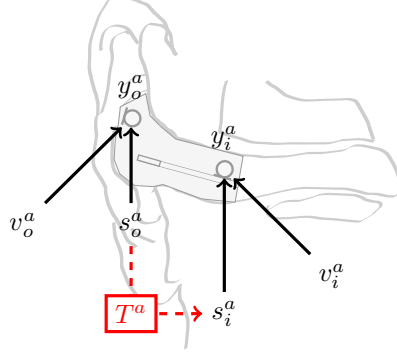


Figure 2.1: The own voice signal model for a hearable with two microphones (outer face, in-ear).

talker-averaged models, by investigating utterance and talker mismatch separately, and by conducting experiments on a larger corpus of hearable recordings.

The paper is structured as follows: In Section 2.2, the own voice signal model is introduced. In Section 2.3, several system identification approaches to model own voice transfer characteristics using time-invariant or time-varying linear filters are presented. In Section 2.4, the performance of these models is evaluated using recorded own voice signals for different conditions.

2.2 Signal model

Figure 2.1 depicts a hearable device equipped with an in-ear microphone and a microphone at the entrance of the (partly) occluded ear canal. The signals at both microphones are denoted by subscripts i and o , respectively. We assume that the hearable is worn by a person (referred to as talker) in a noiseless environment. In the time domain, $s_i^a[n]$ and $s_o^a[n]$ denote the own voice component of talker a at both microphones, where n denotes the discrete-time index. The in-ear microphone signal $y_i^a[n]$ consists of the own voice component and additive noise, i.e.

$$y_i^a[n] = s_i^a[n] + v_i^a[n], \quad (2.1)$$

where the noise component $v_i^a[n]$ consists of unavoidable body-produced noise (e.g., breathing sounds, heartbeats). Similarly, the microphone signal at the entrance of the occluded ear canal $y_o^a[n]$ can be written as

$$y_o^a[n] = s_o^a[n] + v_o^a[n], \quad (2.2)$$

where $v_o^a[n]$ mainly consists of sensor noise. The sensor noise is assumed to be negligible compared to the own voice component in both microphone signals. The own voice components of talker a at the in-ear microphone and the microphone at

the entrance of the occluded ear canal $s_o^a[n]$ are assumed to be related by the own voice transfer characteristics $T^a\{\cdot\}$, i.e.

$$s_i^a[n] = T^a\{s_o^a[n]\}. \quad (2.3)$$

Due to individual anatomical differences of the ear canal [4], these transfer characteristics depend on the talker. In addition, it has been shown that these transfer characteristics depend on the spoken sounds [5, 6] (see also Figure 2.7).

In this paper, we assume that the own voice transfer characteristics $T^a\{\cdot\}$ can be modeled as a *time-varying linear system*, i.e.

$$s_i^a[n] = H^a(q, n) \cdot s_o^a[n], \quad (2.4)$$

with

$$H^a(q, n) = \mathbf{h}^T[n]\mathbf{q}. \quad (2.5)$$

The vector $\mathbf{h}[n]$ denotes a time-varying finite impulse response (FIR) filter with N coefficients,

$$\mathbf{h}[n] = [h_0[n], \quad h_1[n] \quad \dots, \quad h_{N-1}[n]]^T, \quad (2.6)$$

with $\{\cdot\}^T$ the transpose operator, and the vector $\mathbf{q}$ is defined as [28]

$$\mathbf{q} = [1, \quad q^{-1}, \quad \dots, \quad q^{-N+1}]^T, \quad (2.7)$$

with q^{-1} the delay operator. The filtering operation in (2.4) can be approximated in the short-time Fourier transform (STFT) domain as

$$S_i^a(k, l) = H^a(k, l) \cdot S_o^a(k, l), \quad (2.8)$$

where k denotes the frequency bin index, l denotes the time frame index and $H^a(k, l)$ denotes the relative transfer function (RTF) between the microphone at the entrance of the occluded ear canal and the in-ear microphone. Different from (2.4), this approximation is only time-varying between STFT frames and not within a single STFT frame¹.

2.3 Modeling of own voice transfer characteristics

In this section, several methods are presented to model own voice transfer characteristics and subsequently simulated in-ear own voice signals. As outlined in Fig. 2.2, in the *system identification step* the parameters θ of the model $\hat{T}_\theta\{\cdot\}$ are estimated (either in time domain or in frequency domain) based on the signals recorded at the in-ear microphone and the microphone at the entrance of the occluded ear canal. In the *simulation step*, this model can then be used to generate simulated in-ear own voice signals from microphone signals at the entrance of the occluded ear canal, i.e.

$$\hat{s}_i^b[n] = \hat{T}_\theta\{y_o^b[n]\}. \quad (2.9)$$

¹ Circular convolutions effects are also neglected in this approximation, but can be reduced by appropriate windowing.

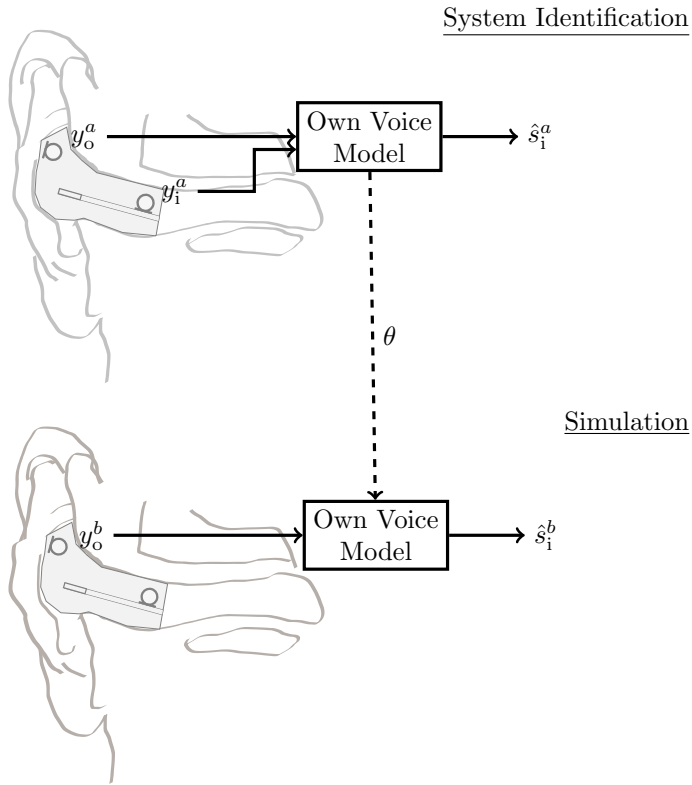


Figure 2.2: Overview of the system identification and simulation steps of the own voice transfer characteristic models.

Both individual models for a specific talker and talker-averaged models will be considered. In Section 2.4 it will be experimentally investigated whether talker-averaging increases robustness to talker mismatch. To estimate the individual model $\hat{T}_\theta^a$ for talker a , recorded microphone signals from talker a are used. This model can then be used to simulate in-ear signals either for the same talker a and the same recorded microphone signals (same talker, same utterance), for different utterances of talker a than used during system identification (utterance mismatch), or for utterances of another talker b (talker mismatch). To estimate the talker-averaged model $\hat{T}_\theta^{\text{avg}}$, recorded microphone signals from several talkers are used.

Sections 2.3.1-2.3.3 consider RTF-based frequency-domain models for the own voice transfer characteristics. In Section 2.3.1, a speech-independent time-invariant model for a specific talker is presented, similarly as in [17]. In Section 2.3.2, a speech-dependent model for a specific talker is proposed, which accounts for the time-varying own voice transfer characteristics by assuming a different RTF for each phoneme. Section 2.3.3 describes how to compute talker-averaged speech-independent and speech-dependent models. Contrary to Sections 2.3.1-2.3.3, in Section 2.3.4 an adaptive filtering-based time-domain model of own voice transfer characteristics is presented, which is utterance-specific.

2.3.1 Speech-independent individual model

If own voice transfer characteristics are assumed to be speech-independent, the individual transfer characteristics of talker a can be modeled as a time-invariant RTF $H^a(k)$ between the microphone at the entrance of the occluded ear canal and the in-ear microphone:

$$\theta_{\text{sp.-indep.}}^a = \left\{ \hat{H}^a(k) \mid k = 1, \dots, K \right\}, \quad (2.10)$$

where K denotes the STFT size. Assuming that the own voice component S_o^a at the entrance of the occluded ear canal and the body-produced noise V_i^a are independent, in the *system identification step* the RTF $\hat{H}^a(k)$ can be estimated using the well-known least squares approach [29], i.e.

$$\hat{H}^a(k) = \arg \min_{H^a(k)} \sum_l |Y_i^a(k, l) - H^a(k) \cdot Y_o^a(k, l)|^2, \quad (2.11)$$

considering all STFT frames of the recorded microphone signals from talker a used for system identification. The least-squares RTF estimate is obtained as

$$\hat{H}^a(k) = \frac{\sum_l Y_i^a(k, l) \cdot Y_o^{a,*}(k, l)}{\sum_l |Y_o^a(k, l)|^2}, \quad (2.12)$$

where $*$ denotes complex conjugation. In the *simulation step*, own voice speech of talker b recorded at the microphone at the entrance of the occluded ear canal is filtered in the STFT domain with the RTF estimate of talker a (where talker a and b can be the same or different), i.e.

$$\hat{S}_i^b(k, l) = \hat{H}^a(k) \cdot Y_o^b(k, l). \quad (2.13)$$

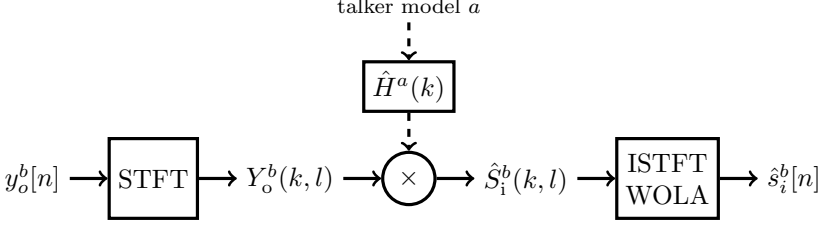


Figure 2.3: Simulation of in-ear own voice signals for talker b using the speech-independent model for talker a .

After applying the inverse STFT, a weighted overlap-add (WOLA) scheme is employed to obtain the time domain signal $\hat{s}_i^b[n]$. Figure 2.3 depicts the signal flow to simulate in-ear own voice signals for talker b using the speech-independent individual model for talker a .

2.3.2 Speech-dependent individual model

Since own voice transfer characteristics likely depend on speech content, we propose to model the transfer characteristics T^a of talker a using a time-varying speech-dependent model. In the *system identification step*, first a frame-wise phoneme annotation $p(l) \in 1, \dots, P$ with P possible phoneme classes is obtained from the microphone signal $y_o^a[n]$ at the entrance of the occluded ear canal using a phoneme recognition system $R\{\cdot\}$:

$$p(l) = R\{y_o^a[n]\}. \quad (2.14)$$

Assuming that the transfer characteristics for each phoneme can be modeled using a (time-invariant) RTF, the RTF for phoneme p' can be estimated from all frames where this phoneme is detected as

$$\hat{H}_{p'}^a(k) = \frac{\sum_{p(l)=p'} Y_i^a(k, l) \cdot Y_o^{a,*}(k, l)}{\sum_{p(l)=p'} |Y_o^a(k, l)|^2}. \quad (2.15)$$

Hence, the speech-dependent model for talker a consists of P RTFs:

$$\theta_{\text{sp.-dep.}}^a = \left\{ \hat{H}_p^a(k) \mid p \in 1, \dots, P, k = 1, \dots, K \right\}. \quad (2.16)$$

In the *simulation step*, first the phoneme sequence $p^b(l)$ is determined on the own voice speech of talker b recorded at the microphone at the entrance of the occluded ear canal. For each frame, the corresponding phoneme-specific RTF $\hat{H}_{p^b(l)}^a(k)$ is selected. In order to prevent discontinuities in the RTFs during phoneme transitions, recursive smoothing with smoothing constant α is applied, i.e.

$$\tilde{H}_{p^b(l)}^a(k) = \alpha \cdot \tilde{H}_{p^b(l-1)}^a(k) + (1 - \alpha) \cdot \hat{H}_{p^b(l)}^a(k). \quad (2.17)$$

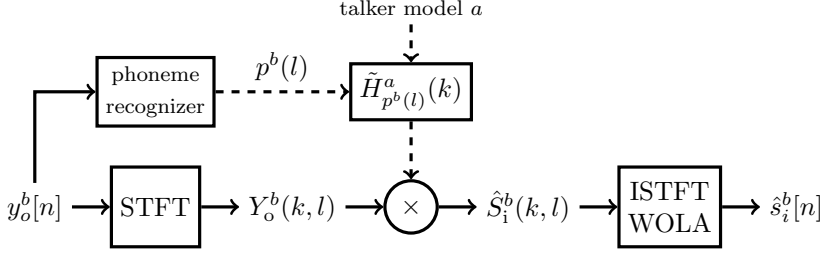


Figure 2.4: Simulation of in-ear own voice signals for talker b using the proposed speech-dependent model for talker a .

The smoothed RTF $\tilde{H}_{p^b(l)}^a(k)$ is then used to simulate the own voice of talker b at the in-ear microphone:

$$\hat{S}_i^b(k, l) = \tilde{H}_{p^b(l)}^a(k) \cdot Y_o^b(k, l). \quad (2.18)$$

Similarly to the speech-independent model, a WOLA scheme is employed to obtain the time-domain signal $\hat{s}_i^b[n]$. Figure 2.4 depicts the signal flow to simulate in-ear own voice signals for talker b using the speech-dependent model for talker a . Due to the phoneme recognition system for frame-wise phoneme-specific RTF selection, we expect that the proposed speech-dependent model is able to simulate in-ear signals more accurately than the speech-independent model, also for utterances not used during system identification. In addition, it should be realized that unlike the speech-independent model, the speech-dependent model also accounts for speech pauses by modeling them as a separate phoneme.

2.3.3 Talker-averaged models

Since individual models may not generalize well to different talkers, we also consider talker-averaged speech-independent and speech-dependent models. In the *system identification step*, talker-averaged models are obtained by considering all STFT frames of the recorded microphone signals of all utterances from all talkers except talker b (leave-one-out-paradigm) for system identification. The RTFs of the speech-independent talker-averaged model are hence computed as

$$\hat{H}^{\text{avg}}(k) = \frac{\sum_{a \neq b} \sum_l Y_i^a(k, l) \cdot Y_o^{a,*}(k, l)}{\sum_{a \neq b} \sum_l |Y_o^a(k, l)|^2}, \quad (2.19)$$

while the RTFs of the speech-dependent talker-averaged model for phoneme p' are computed as

$$\hat{H}_{p'}^{\text{avg}}(k) = \frac{\sum_{a \neq b} \sum_{p(l)=p'} Y_i^a(k, l) \cdot Y_o^{a,*}(k, l)}{\sum_{a \neq b} \sum_{p(l)=p'} |Y_o^a(k, l)|^2}. \quad (2.20)$$

The *simulation step* for the talker-averaged models is similar as for the individual models, where for the speech-independent model $\hat{H}^{\text{avg}}(k)$ is used instead of $\hat{H}^a(k)$ and for the speech-dependent model $\hat{H}_{p'}^{\text{avg}}(k)$ is used instead of $\hat{H}_{p'}^a(k)$.

2.3.4 Adaptive filtering-based model

As an alternative to the time-varying speech-dependent model in Section 2.3.2, in this section we consider a time-domain adaptive filter to model the time-varying transfer path between the microphone at the entrance of the occluded ear canal and the in-ear microphone. The signal flow is illustrated in Figure 2.5. In the *system identification step*, the FIR filter $\hat{\mathbf{h}}^a[n]$ with N coefficients is adapted based on recorded microphone signals of an utterance of talker a . The adaptive filter aims at minimizing the error between the in-ear microphone signal $y_i^a[n]$ and the estimated in-ear own voice signal

$$\hat{s}_i^a[n] = \hat{H}^a(q, n) \cdot y_o^a[n] = \left(\hat{\mathbf{h}}^a[n] \right)^T \mathbf{y}_o^a[n], \quad (2.21)$$

with

$$\mathbf{y}_o^a[n] = \left[y_o^a[n], y_o^a[n-1], \dots, y_o^a[n-N+1] \right]^T. \quad (2.22)$$

For adapting the filter the well-known normalized least mean squares (NLMS) algorithm is used [26], i.e. the filter coefficients are recursively updated as

$$\hat{\mathbf{h}}^a[n+1] = \hat{\mathbf{h}}^a[n] + \frac{\mu}{\epsilon + (\mathbf{y}_o^a[n])^T \mathbf{y}_o^a[n]} \mathbf{y}_o^a[n] \left(y_i^a[n] - \left(\hat{\mathbf{h}}^a[n] \right)^T \mathbf{y}_o^a[n] \right), \quad (2.23)$$

where μ denotes the step size and ϵ is a small regularization constant. The model parameters of the adaptive filtering-based model are

$$\theta_{\text{adapt.}}^a = \{ \hat{\mathbf{h}}^a[n], n = 1, \dots \}. \quad (2.24)$$

Since this model implicitly depends on a specific utterance, it should be noted that it is not possible to obtain a talker-averaged model by following a similar procedure as described in the previous section.

In the *simulation step*, the simulated in-ear own voice signal of talker b is computed as

$$\hat{s}_i^b[n] = \left(\hat{\mathbf{h}}^a[n] \right)^T \mathbf{y}_o^b[n]. \quad (2.25)$$

In case of utterance mismatch (both for the same talker and for a different talker), the filter is applied to a different input signal than used during adaptation which likely results in estimation errors.

2.4 Experimental evaluation

In this section, the own voice transfer characteristic models discussed in Section 2.3 are evaluated in terms of their accuracy in simulating in-ear own voice signals for different conditions. In Section 2.4.1, the data used in the evaluation and the experimental conditions are described. In Section 2.4.2, the simulation parameters are defined. In Section 2.4.3, examples of simulated in-ear own voice signals and estimated RTFs are presented for all considered RTF-based models. In Sections 2.4.4-2.4.6, experimental results are presented and discussed for three conditions: matched condition (same talker, same utterance), utterance mismatch and talker mismatch.

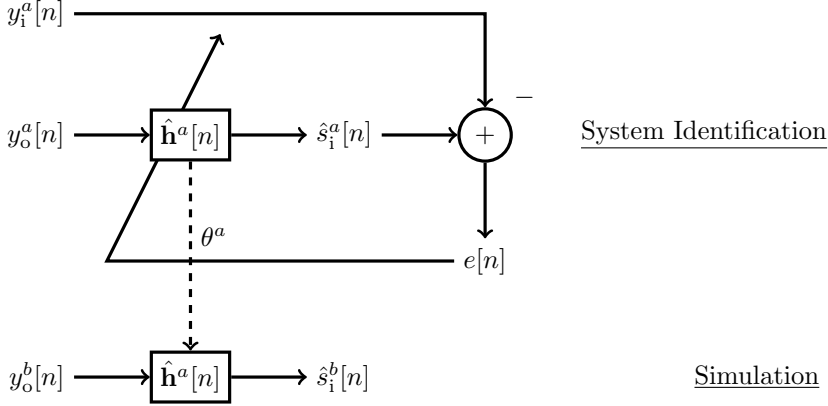


Figure 2.5: The adaptive filtering scheme utilized for estimating in-ear speech signals. The filter coefficients are transferred from system identification to simulation directly after each sample-wise adaptation step.

2.4.1 Recording setup and experimental conditions

For identifying and evaluating the own voice transfer characteristic models, we recorded a dataset of own voice speech from 18 native German talkers (5 female, 13 male), with approximately 25 to 30 minutes of recorded own voice signals per talker. The hearable device used for recording is the closed-vent variant of the one-size-fits-all Hearpiece [25]. The Hearpiece *concha* microphone of the device was selected as the microphone at the outer face of the occluded ear canal. Talkers were excluded if insertion of the hearable was not possible, or if bad fittings with insufficient attenuation of external sounds were detected (by measuring a transfer function from an external loudspeaker between the concha and in-ear microphone). For each talker, 306 pre-determined sentences were recorded: The Marburg and Berlin sentences [30], each consisting of 100 sentences, 100 common everyday German sentences for language learners [31], and the German version of the well-known text *The North Wind and the Sun*, consisting of 6 sentences. Recordings were conducted in a sound-proof listening booth using a Behringer UMC1820 audio interface. Before the recordings started, informed consent was obtained from all talkers. The recorded dataset is publicly available on Zenodo [32]. During system identification, model parameters were estimated on 150 sentences uttered by each talker. During simulation, in-ear own voice signals are generated from the recorded microphone signals at the outer face of the Hearpiece and evaluated per utterance.

Three different simulation conditions are investigated:

2.4.1.1 Same talker, same utterance (matched condition)

In this condition, the individual RTF-based models and the adaptive filtering-based model are evaluated exactly the same utterances of the same talker ($a = b$) as

considered during model estimation. For the adaptive filtering-based model, this means that the same signal $y_o^b[n] = y_o^a[n]$ is used during simulation as during identification (see Figure 2.5), such that the simulated in-ear signal $\hat{s}_i^b[n]$ is equal to the output of the adaptive filter $\hat{s}_i^a[n]$. Talker-averaged models are not considered in this condition.

2.4.1.2 *Same talker, utterance mismatch*

In this condition, the individual RTF-based models and the adaptive filtering-based model are evaluated on speech of the same talker ($a = b$) as considered during model estimation. In order to investigate the generalization ability of the models for the same talker, evaluation is performed on the 156 sentences not used to estimate the models. For the adaptive filtering-based model, the length of the signals used during simulation and identification is matched, either by cutting or concatenating the signals used during model estimation with other signals from the same talker. Talker-averaged models are not considered in this condition.

2.4.1.3 *Talker mismatch*

The generalization ability of models to unseen talkers is investigated by estimating speech of talker b using models estimated on a different talker ($a \neq b$). For each utterance, a random talker a is assigned to talker b . In this condition, there is also an implicit utterance mismatch because the same sentence uttered by different talkers most likely has differences with respect to speed, frequency content, pronunciation and other speech attributes. Talker-averaged models are considered in this condition only. For each talker b , a talker-averaged model is computed from utterances of the remaining 17 talkers. Evaluation is performed on the 156 sentences not used to estimate the models. In all three conditions, log-spectral distance (LSD) [33] and mel-cepstral distance (MCD) [34] between the recorded in-ear signals $y_i^b[n]$ and the simulated in-ear signals $\hat{s}_i^b[n]$ are used as evaluation metrics. For both metrics, a lower value indicates a more accurate estimate. Since perceptual metrics such as PESQ [35] were found not to correlate well with subjective ratings of body-conducted own voice signals [36], such metrics are not considered in this study.

2.4.2 *Simulation parameters*

The experiments were carried out at a sampling frequency of 5 kHz, since above 2.5 kHz the in-ear microphone signals hardly contain any body-conducted speech for the considered hearable device. Model-specific parameters were set empirically based on preliminary experiments. For the RTF-based models, an STFT framework with a frame length of $K = 128$ (corresponding to 25.6 ms) and an overlap of 50 % was used, where a square-root Hann window was utilized both as analysis and synthesis window. For the speech-dependent models, a smoothing parameter of $\alpha = 0.8$ was used in (2.17), corresponding to an effective smoothing time of 64 ms. The used phoneme recognition system was trained on German speech and $P = 62$ phoneme classes. For the adaptive filtering-based model, the filter length

was set to $N = 128$, and a step size parameter $\mu = 0.5$ and regularization constant $\epsilon = 10^{-6}$ were used in (2.23). The filter coefficients were initialized as zeroes. For all methods, no voice activity detection was employed so that utterances may contain short pauses.

2.4.3 Example spectrograms and RTFs

For the RTF-based models, this section presents examples of simulated in-ear own voice signals, spectrograms and estimated RTFs. For the matched condition (same talker, same utterance), Figure 2.6 for a specific utterance (the beginning of *The North Wind and the Sun*) of talker 2 (male). The shown spectrograms are the spectrograms of the microphone signal at the entrance of the occluded ear canal and the in-ear microphone signal as well as the in-ear own voice signals simulated with the speech-independent models and the proposed speech-dependent models (individual and talker-average)². While it can be observed that the speech-independent models estimate the in-ear microphone signal rather well in the frequency region below 500 Hz, they clearly underestimate own voice components for higher frequencies. On the other hand, the speech-dependent models are able to estimate the in-ear microphone signal more accurately at higher frequencies, although deviations are visible above 1 kHz. The estimates of individual and talker-averaged models are very similar for both the speech-independent and speech-dependent models for this example. It should be noted that the low-frequency body-produced noise in the in-ear microphone signal is not present in all simulated in-ear own voice signals. For the same utterance as in Figure 2.6, Figure 2.7 depicts the time-domain own voice signal recorded at the entrance of the occluded ear canal with its phoneme annotation, and the magnitude of the phoneme-specific individual RTFs, estimated using (2.15). Different from other experiments, these RTFs were estimated with a sampling frequency of 16 kHz and an STFT size of $N = 256$ to show the high-frequency region as well. It can be seen that for different phonemes, the RTFs differ a lot in the low-frequency region below 2.5 kHz, while above 2.5 kHz the RTFs are very similar.

To compare the RTF-based models, Figure 2.8 depicts the estimated RTF magnitudes for the speech-independent models (top subplot) and the speech-dependent models for two selected phonemes (middle and bottom subplot), considering all talkers in the experiments. The individual RTFs are represented by shaded regions and the talker-averaged RTFs as solid lines. Different from the talker-averaged RTFs used in the talker mismatch condition (leave-one-out-paradigm), averages here are computed over all 18 talkers. For the speech-independent RTFs, it can be observed that for most talkers the low frequency region below approximately 600 Hz is amplified at the in-ear microphone relative to the microphone at the entrance of the occluded ear canal, whereas the frequency region above approximately 1.5 kHz is attenuated. While half of the estimated RTFs (i.e., between the quartiles Q1 and Q3) are very similar in magnitude, for some talkers there appear to be larger deviations

² Audio examples corresponding to the spectrograms are available online at <https://doi.org/10.5281/zenodo.11371976> [37].

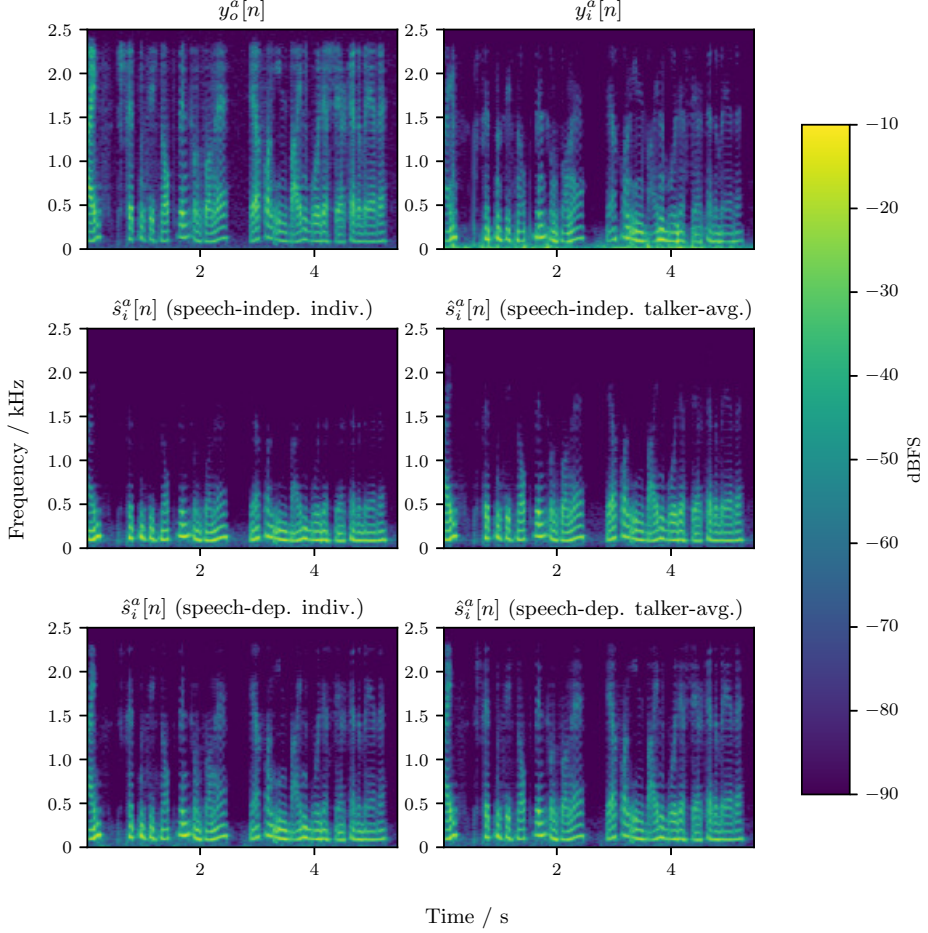


Figure 2.6: Example spectrograms for the same talker, same utterance condition: recorded own voice signal of talker 2 at the entrance of the occluded ear canal (top left) and recorded in-ear own voice signal (top right) of talker 2, and the simulated in-ear own voice signals estimated by the speech-independent individual (middle left) and speech-independent talker-averaged (middle right), and the speech-dependent individual (bottom left) and speech-dependent talker-averaged (bottom right) models.

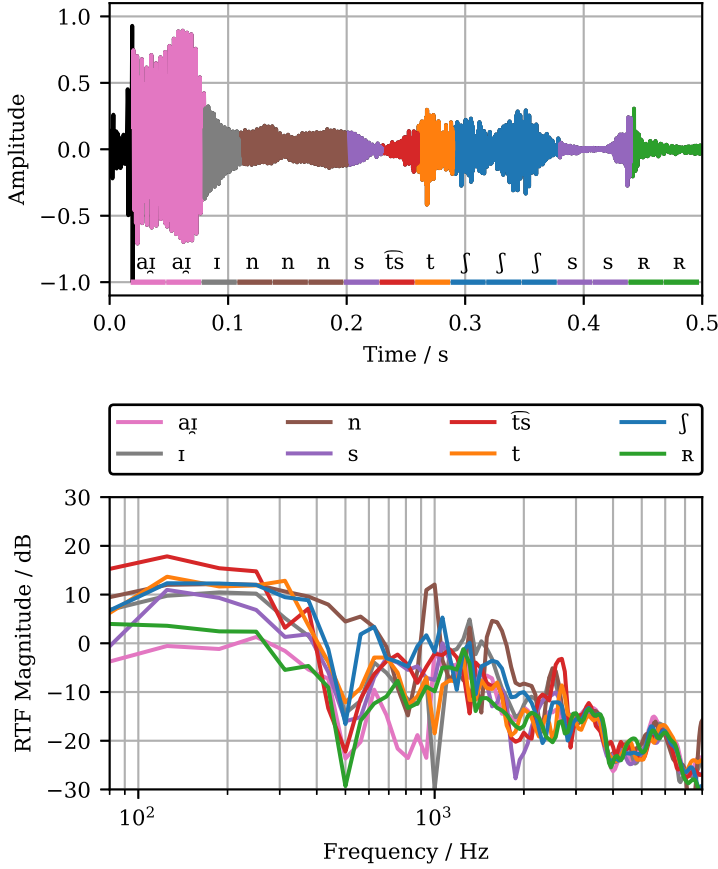


Figure 2.7: Example own voice signal of talker 2 recorded at the entrance of the occluded ear canal with phoneme annotation (top) and magnitude of phoneme-specific individual relative transfer functions (bottom) estimated on all utterances of this talker (speech-dependent individual model). Only RTF magnitudes of phonemes appearing in the depicted utterance are shown.

from the talker-averaged RTF magnitude. For the phoneme-specific RTFs shown in the middle and lower subplot, similar tendencies in terms of inter-individual variance can be observed. However, it can be observed that the phoneme-specific talker-averaged RTFs differ from the speech-independent talker-averaged RTFs. In particular, for the phoneme /3/ the magnitude is considerably higher than the magnitude of the speech-independent talker-averaged RTF in the frequency region between 500 Hz to 1.5 kHz and above 2 kHz for the majority of talkers. In contrast, for the phoneme /o/ the RTF magnitudes are lower than the magnitude of the speech-independent talker-averaged RTF especially in the low frequency region.

2.4.4 *Same talker, same utterance*

For the matched condition (same talker, same utterance), Figure 2.9 shows the LSD and MCD scores between the recorded in-ear signals and the simulated in-ear signals for the speech-independent and speech-dependent individual RTF-based models and the adaptive filtering-based model. It can be observed that both metrics are much lower for the speech-dependent individual model and the adaptive filtering-based model than for the speech-independent individual model. These results demonstrate that in-ear own voice signals can be simulated more accurately when time-varying or speech-dependent transfer characteristics are accounted for. In addition, the speech-dependent individual model performs nearly as well as the adaptive filtering-based model, where it should be realized that for the matched condition the (utterance-specific) adaptive filter can be considered as the optimal time-varying filter. This indicates that the proposed phoneme-specific RTF-based model is able to accurately model time-varying behavior of own voice transfer characteristics. It can be noted that even in the matched condition, none of the considered methods is able to perfectly simulate the recorded in-ear own voice signals. This can be explained by the fact that the considered methods are not able to account for body-produced noise (see Figure 2.6) and possible non-linear effects, which are however assumed to be small.

2.4.5 *Same talker, utterance mismatch*

For the same models as in the previous section, Figure 2.10 shows the LSD and MCD score for the utterance mismatch condition (same talker, utterance mismatch). The results for the speech-dependent and speech-independent individual models are very similar in the matched condition (see Figure 2.9), indicating that both models generalize well to other utterances of the same talker. For the adaptive filtering-based model, on the other hand, the LSD and MCD scores are much larger than for the matched condition, showing that the utterance-specific adaptive filtering-based method (expectedly) does not generalize well to other utterances.

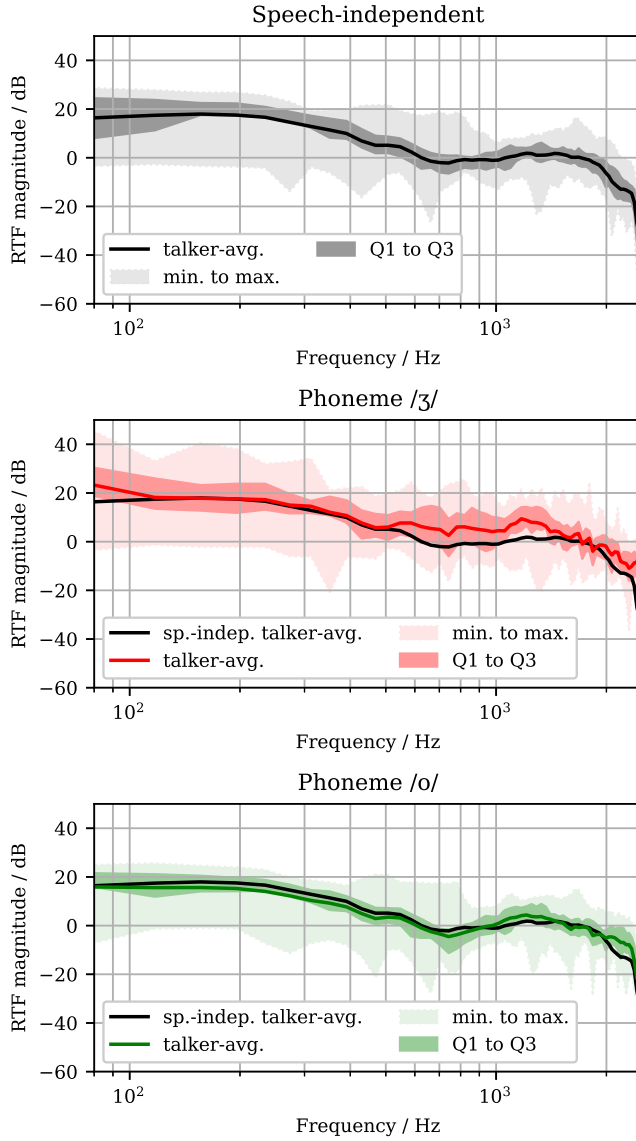


Figure 2.8: Relative transfer functions estimated for the speech-independent individual and talker-averaged models (top) and for two phonemes with the speech-dependent models (middle and bottom). Values between the quartiles Q1 and Q3 and between the minimum and maximum values of the individual models are indicated by shaded regions. Talker-averaged relative transfer functions over all talkers are shown as solid black lines.

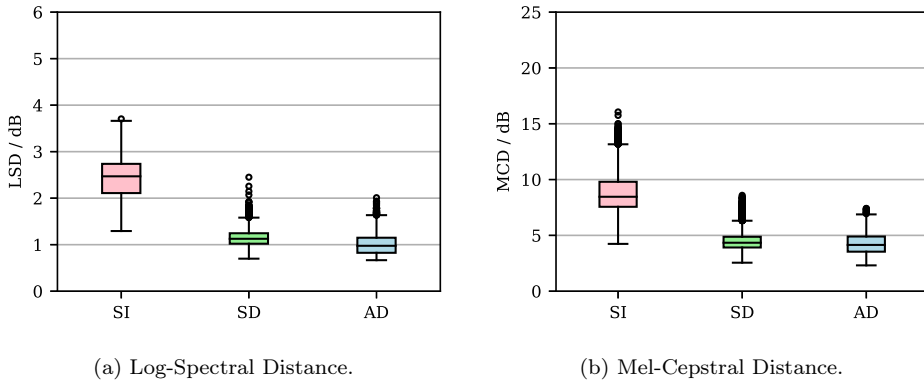


Figure 2.9: Results for the *same talker, same utterance* condition with speech-independent (SI), speech-dependent (SD) and adaptive filtering-based (AD) models. Note that the y-axis limits of both subfigures are different.

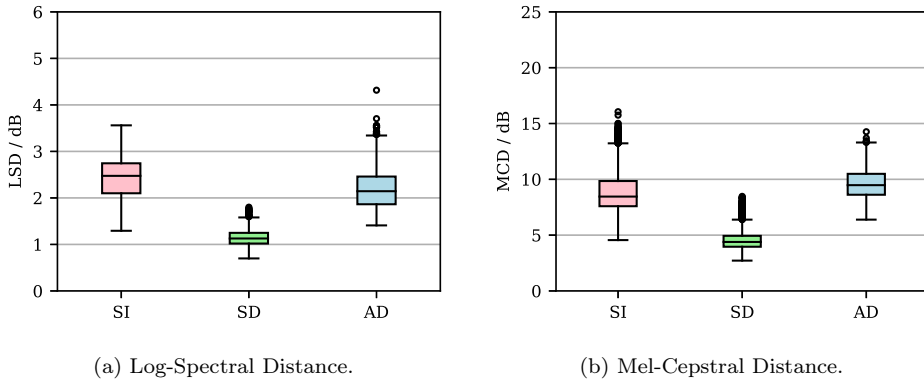


Figure 2.10: Results for the *same talker, utterance mismatch* condition with speech-independent (SI), speech-dependent (SD) and adaptive filtering-based (AD) models. Note that the y-axis limits of both subfigures are different.

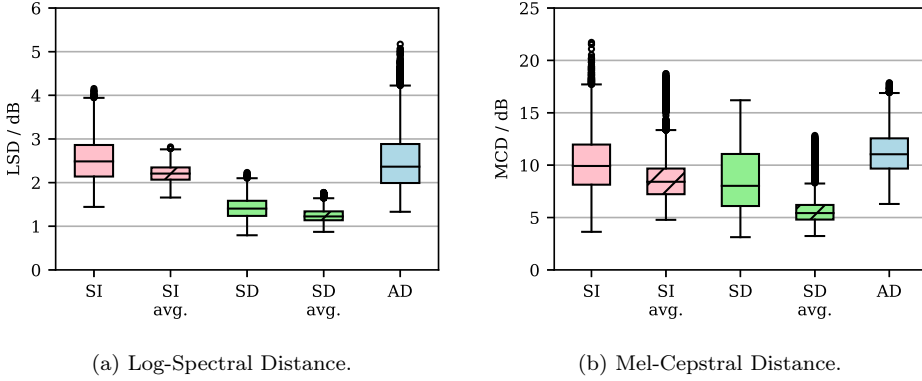


Figure 2.11: Results for the *talker mismatch* condition with speech-independent (SI), speech-dependent (SD) and adaptive filtering-based (AD) models, using individual and talker-averaged (avg.) versions. Note that the y-axis limits of both subfigures are different.

2.4.6 *Talker mismatch*

For the talker mismatch condition, Figure 2.11 shows the LSD and MCD scores for the speech-independent and speech-dependent models (both individual and talker-averaged) and the adaptive filtering-based model. It can be clearly observed that the speech-dependent models outperform the speech-independent models and the adaptive filtering-based model, where the best performance in terms of both metrics is achieved by the speech-dependent talker-averaged model. This indicates that the speech-dependent talker-averaged model has the best generalization ability to unseen talkers. Comparing the results in Figure 2.10 and Figure 2.11, it can be observed that the LSD and MCD scores of the speech-dependent individual model are larger under talker mismatch. Especially the large variance of the MCD score is noticeable. Since this effect does not occur in the other conditions, it is likely a consequence of talker mismatch.

2.5 Discussion

The experiments in Section 2.4 investigated models of own voice transfer characteristics for simulating in-ear own voice signals. While adaptive filters cannot be used in practice since they are utterance-specific, the proposed speech-dependent RTF-based models are able to generalize to unseen utterances. In case of talker mismatch, the speech-dependent talker-averaged model was more robust than the speech-dependent individual model.

2.5.1 *Limitations*

It needs to be realized that due to the usage of a phoneme recognition system, the proposed speech-dependent models exhibit several limitations: First, since the considered phoneme recognition system has been trained with German speech only, the speech-dependent models may not generalize well to other languages, or may require a phoneme recognition system matching these languages. Second, the phoneme recognition system, which is based on a speech recognition system, computes its phoneme annotation when entire words are recognized. This leads to a variable processing delay, typically in the range of several hundred milliseconds to one second. With this phoneme recognition system, the proposed models cannot be used for real-time, low-latency applications, so that a different phoneme recognition system with a lower processing delay may be better suited for these applications. Third, the models are limited to the specific device used to obtain the recorded signals for model estimation. Applying the models to simulate in-ear own voice signals for other devices (e.g., over-ear headphones) would require estimation of RTFs from own voice signals recorded with those devices. Finally, the phoneme-dependent RTFs in the proposed models are estimated for discrete phonemes, and phoneme transitions are handled by temporal smoothing (see Section 2.3.2). However, this approximation may not accurately reflect the actual mouth movements that occur between uttering two phonemes.

2.5.2 *Comparison to previous research*

While previous research has addressed simulating in-ear own voice signals, the influence of speech-dependent changes has not been investigated specifically. Earlier studies either focus on speech-independent or black-box DNN models, or are not concerned with simulating in-ear own voice signals. In [7], occlusion effect level differences were modeled for several phonemes using a linear regression model that relates the phoneme formant frequencies to the amount of occlusion in the relevant frequency region below 500 Hz. However, the model in [7] does not allow for the simulation of new in-ear own voice signals. In [24], a DNN model was proposed to convert air-conducted to bone-conducted speech, accounting for individual differences between talkers based on a speaker identification system. While the model was able to generalize to different talkers than those used during training, the role of speech-dependent changes was not investigated. Recently, several DNN-based approaches have been proposed for own voice reconstruction (i.e. reconstruction of own voice speech from hearable microphones) either only using an in-ear microphone or a body-conduction sensor without considering environmental noise [17, 18, 38], or using both a body-conduction sensor and a microphone at the outer face of a hearable while considering environmental noise [39]. To simulate own voice signals for training, these approaches introduce random variations, either by using several RTFs per talker [17, 38] or adding random values to the RTFs [18, 39]. However, the accuracy of these approaches for simulating in-ear own voice signals has not

been investigated, and speech-dependent changes were not accounted for in the simulation.

2.5.3 Applications

Due to their robustness to utterance and talker mismatch, the proposed speech-dependent models may be used, e.g., to simulate in-ear own voice signals as training data for DNN-based algorithms aiming at joint bandwidth extension, equalization, and noise reduction of own voice signals recorded at an in-ear or body-conduction microphone. In this application, a large amount of own voice signals is typically required to train DNNs. The proposed models may be beneficial for these applications, as they may be used to simulate in-ear own voice signals from broadband speech signals. Speech-independent models have already been used for this purpose in [17, 38, 39]. Since in-ear or body-conduction microphones are also beneficial for speech recognition systems (see e.g., [40]), the proposed speech-dependent models could be applied to training an own voice speech recognition system by simulating training data.

2.6 Conclusion

In this paper, speech-dependent models of own voice transfer characteristics in hearables have been proposed. The models can be utilized to estimate own voice signals at an in-ear microphone. In particular, the proposed models take into account time-varying speech-dependent behavior and inter-individual differences between talkers. To estimate in-ear own voice signals from broadband speech using the proposed speech-dependent models, phoneme-specific RTFs are used. The influence of utterance and talker mismatch on the estimation accuracy of in-ear own voice signals has been investigated in an experimental evaluation. Results show that using a speech-dependent model is beneficial compared to using a speech-independent model. Although the adaptive filtering-based approach is able to model the speech-dependency of the own voice transfer characteristics well in the matched condition, it completely fails when considering utterance and talker mismatch. However, the proposed individual speech-dependent models are able to generalize to different utterances of the same talker. Talker-averaged models were shown to generalize better to different talkers than individual models. Future work will investigate the usage of the proposed models for simulating in-ear signals to train own voice reconstruction algorithms based on supervised learning.

Author declarations

Conflict of interest

The authors declare no conflict of interest.

Data privacy management

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

Data availability statement

The research data associated with this article are available in Zenodo, under the reference [32]. Supplemental material (listening examples) is available in Zenodo, under the reference [37].

Acknowledgments

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. The authors wish to thank the talkers for their participation in the recordings.

References

- [1] R. E. Bouserhal, A. Bernier, and J. Voix, “An in-ear speech database in varying conditions of the audio-phonation loop,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 1069–1077, Feb. 2019.
- [2] M. Ø. Hansen, “Occlusion effects Part I and II,” PhD thesis, Department of Acoustic Technology, Technical University of Denmark, 1998.
- [3] S. Stenfelt and S. Reinfeldt, “A model of the occlusion effect with bone-conducted stimulation,” *International Journal of Audiology*, vol. 46, no. 10, pp. 595–608, Jan. 2007.
- [4] S. Vogl and M. Blau, “Individualized prediction of the sound pressure at the eardrum for an earpiece with integrated receivers and microphones,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 917–930, Feb. 2019.
- [5] S. Reinfeldt, P. Östli, B. Håkansson, and S. Stenfelt, “Hearing one’s own voice during phoneme vocalization - Transmission by air and bone conduction,” *J. Acoust. Soc. Am.*, vol. 128, no. 2, pp. 751–762, Aug. 2010.
- [6] H. Saint-Gaudens, H. Nélisse, F. Sgard, and O. Doutres, “Towards a practical methodology for assessment of the objective occlusion effect induced by earplugs,” *J. Acoust. Soc. Am.*, vol. 151, no. 6, pp. 4086–4100, Jun. 2022.

- [7] T. Zurbrügg, A. Stirnemann, M. Kuster, and H. Lissek, “Investigations on the physical factors influencing the ear canal occlusion effect caused by hearing aids,” *Acta Acustica united with Acustica*, vol. 100, no. 3, pp. 527–536, May 2014.
- [8] J. Richard, V. Zimpfer, and S. Roth, “Effect of bone conduction microphone location and mouth opening on transfer function between oral cavity sound pressure and skin acceleration,” in *Proc. Convention of the European Acoustics Association (Forum Acusticum)*, Turin, Italy, Sep. 2023.
- [9] C. Pörschmann, “Influences of bone conduction and air conduction on the sound of one’s own voice,” *Acta Acustica united with Acustica*, vol. 86, no. 6, pp. 1038–1045, Nov. 2000.
- [10] M. K. Brummund, F. Sgard, Y. Petit, and F. Laville, “Three-dimensional finite element modeling of the human external ear: Simulation study of the bone conduction occlusion effect,” *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1433–1444, Mar. 2014.
- [11] S. Liebich, J. Fabry, P. Jax, and P. Vary, “Signal Processing Challenges for Active Noise Cancellation Headphones,” in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 11–15.
- [12] P. Rivera Benois, R. Roden, M. Blau, and S. Doclo, “Optimization of a Fixed Virtual Sensing Feedback ANC Controller For In-Ear Headphones with Multiple Loudspeakers,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 8717–8721.
- [13] T. Zurbrügg, “The Occlusion Effect - Measurements, Simulations and Countermeasures,” in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 26–30.
- [14] S. Liebich and P. Vary, “Occlusion Effect Cancellation in Headphones and Hearing Devices—The Sister of Active Noise Cancellation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 35–48, 2022.
- [15] R. E. Bouserhal, T. H. Falk, and J. Voix, “In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension,” *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 1321–1331, Mar. 2017.
- [16] H. Wang, X. Zhang, and D. Wang, “Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 3134–3143, 2022.
- [17] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Training strategies for own voice reconstruction in hearing protection devices using an in-ear microphone,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022.
- [18] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, “Configurable EBEN: Extreme bandwidth extension network to enhance body-conducted speech capture,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 3499–3512, 2023.

- [19] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Multi-microphone noise data augmentation for DNN-based own voice reconstruction for hearables in noisy environments,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 416–420.
- [20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5206–5210.
- [21] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5220–5224.
- [22] W. He, P. Motlicek, and J.-M. Odobez, “Neural Network Adaptation and Data Augmentation for Multi-Speaker Direction-of-Arrival Estimation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 1303–1317, 2021.
- [23] P. Srivastava, A. Deleforge, and E. Vincent, “Realistic Sources, Receivers and Walls Improve The Generalisability of Virtually-Supervised Blind Acoustic Parameter Estimators,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022.
- [24] M. Pucher and T. Woltron, “Conversion of Airborne to Bone-Conducted Speech with Deep Neural Networks,” in *Proc. Interspeech*, Brno, Czechia, Aug. 2021, pp. 1–5.
- [25] F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J.-H. Bach, J. Wellmann, and B. Kollmeier, “A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research,” in *Proc. AES International Conference on Headphone Technology*, San Francisco, USA, Aug. 2019.
- [26] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Prentice Hall, 1996.
- [27] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Speech-dependent modeling of own voice transfer characteristics for in-ear microphones in hearables,” in *Proc. Convention of the European Acoustics Association (Forum Acusticum)*, Turin, Italy, Sep. 2023, pp. 1899–1902.
- [28] L. Ljung, “System identification,” in *Signal Analysis and Prediction*, Springer, 1998, pp. 163–173.
- [29] Y. Avargel and I. Cohen, “On multiplicative transfer function approximation in the short-time Fourier transform domain,” *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, May 2007.
- [30] A. P. Simpson, K. J. Kohler, and T. Rettstadt, “The Kiel corpus of read/spontaneous speech: Acoustic data base, processing tools, and analysis results,” in *Arbeitsberichte Institut Für Phonetik Und Digitale Sprachverarbeitung Universität Kiel*, vol. 32, IPDS, Nov. 1997, pp. 243–247.

- [31] A. Neustein, *100 Sätze reichen für ein ganzes Leben (Blog-post)*, <https://deutschlernerblog.de/100-saetze-reichen-fuer-ein-ganzes-leben/>, Aug. 2019. Accessed: Jan. 24, 2023.
- [32] M. Ohlenbusch, C. Rollwage, and S. Doclo, *German own voice recordings with hearable microphones*, Mar. 2024.
- [33] A. Gray and J. Markel, “Distance measures for speech processing,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.
- [34] R. F. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, Victoria, BC, Canada, May 1993, pp. 125–128.
- [35] 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,” *International Telecommunications Union*, Feb. 2001.
- [36] J. Richard, V. Zimpfer, and S. Roth, “Comparison of objective and subjective methods for evaluating speech quality and intelligibility recorded through bone conduction and in-ear microphones,” *Applied Acoustics*, vol. 211, Aug. 2023.
- [37] M. Ohlenbusch, C. Rollwage, and S. Doclo, *Modeling of speech-dependent own voice transfer characteristics for hearables with in-ear microphones: Audio examples*, Zenodo, Mar. 2024.
- [38] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, “Speaker adaptation for enhancement of bone-conducted speech,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 10 456–10 460.
- [39] L. He, H. Hou, S. Shi, X. Shuai, and Z. Yan, “Towards bone-conducted vibration speech enhancement on head-mounted wearables,” in *Proc. Annual International Conference on Mobile Systems, Applications and Services*, New York, USA, Jun. 2023, pp. 14–27.
- [40] M. Wang, J. Chen, X.-L. Zhang, and S. Rahardja, “End-to-end multi-modal speech recognition on an air and bone conducted speech corpus,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 513–524, 2023.

SPEECH-DEPENDENT DATA AUGMENTATION FOR OWN VOICE RECONSTRUCTION WITH HEARABLE MICROPHONES IN NOISY ENVIRONMENTS

This chapter is identical in content to the publication: M. Ohlenbusch, C. Rollwage, and S. Doclo, “Speech-dependent data augmentation for own voice reconstruction with hearable microphones in noisy environments,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, no. 32, 2025. DOI: [10.1186/s13636-025-00418-1](https://doi.org/10.1186/s13636-025-00418-1).

Authors	Author's contribution							
	A	B	C	D	E	F	G	H
Mattes Ohlenbusch	✗	✗	✗	✗	✗	✗	✗	✗
Christian Rollwage	✗			✗		✗	✗	✗
Simon Doclo	✗			✗		✗	✗	✗

A - Substantial contributions to the conception or design of the work

B - Acquisition of the data

C - Analysis of the data

D - Interpretation of the data

E - Drafting the work

F - Revising the work critically

G - Final approval of the version to be published

H - Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Abstract

Hearable devices, equipped with one or more microphones, can be used to capture the user’s own voice in noisy environments. In such environments, an own voice reconstruction (OVR) system is needed to enhance the quality and intelligibility of the recorded own voice. In this work, we aim to estimate clean broadband speech from a microphone at the outer face of the hearable and an in-ear microphone, which captures the own voice at a higher signal-to-noise ratio than the outer microphone, but with a limited bandwidth and additive body-produced noise. Training a supervised deep learning-based OVR system requires a substantial amount of own voice signals as training data. Such training data can be collected by recording many utterances from different talkers wearing the hearable, which is costly, or generated by augmenting existing clean speech datasets. In this paper, we investigate several data augmentation techniques to simulate a large amount of in-ear own voice signals from a limited amount of recorded own voice signals. More specifically, we consider different models for the own voice transfer characteristics between the outer microphone and the in-ear microphone, ranging from a fixed talker-averaged relative transfer function to a phoneme-dependent individual model. We investigate the influence of the amount of recorded own voice signals on the performance of an OVR system based on the FT-JNF architecture, either by directly using the recorded signals for training or by using the recorded signals to generate augmented data for training (with and without fine-tuning with recorded signals). Experimental results show that training using the proposed speech-dependent individual data augmentation technique and additional fine-tuning with recorded signals yields the best performance in terms of objective metrics, even when only few recorded own voice signals are available.

3.1 Introduction

Speech communication is often impaired in noisy environments, such as busy traffic areas or industrial manufacturing sites. In such environments, hearable devices with integrated microphones can be used to improve communication, e.g., by capturing and transmitting the user’s own voice to a mobile phone or another hearable [1, 2]. In addition to applications in person-to-person communication, own voice pickup can also improve automatic speech recognition [3], enabling more robust control of production machines [4], hearing aids [5], and using voice assistants [6]. In this paper, we consider a hearable equipped with two microphones: a microphone at the outer face (outer microphone) and a microphone inside the partly occluded ear canal (in-ear microphone). Since the hearable occluding the ear canal attenuates environmental noise, the in-ear microphone may be particularly beneficial, since it captures the user’s own voice at a higher signal-to-noise ratio than the outer microphone. However, compared to own voice recorded at the outer microphone, own voice recorded at the in-ear microphone is subject to amplification at low frequencies (below approximately 1 kHz) and strong attenuation at higher frequencies (above approximately 2 kHz), resulting in a limited bandwidth and poor signal quality [7]. In addition, the in-ear microphone also records body-produced noise, such as respiratory and heart sounds [8]. It has been shown that the ratio between the airborne and body-conducted components of own voice recorded at an in-ear microphone depends on the phonetic content [9, 10] (e.g., due to mouth movements or place of excitation changes), and the individual user [11, 12]. In this paper, the relationship between own voice recorded at the outer microphone and the in-ear microphone is referred to as own voice transfer characteristics. Several models of own voice transfer characteristics have been presented in the literature. In [13–15] the own voice transfer characteristics are modeled using a time-invariant relative transfer function, whereas in [16] the own voice transfer characteristics are modeled using a relative transfer function for each phoneme, leading to a (time-varying) speech-dependent model.

Since neither the quality of the outer microphone nor the in-ear microphone is sufficient, own voice reconstruction (OVR) algorithms have been proposed that aim at estimating clean broadband speech from the (noisy) outer microphone signal and/or the (band-limited) in-ear microphone signal. Classical signal processing approaches for reconstructing own voice using body-conduction microphones¹ are based on, e.g., equalization filter design [17], linear prediction analysis and synthesis [18], or statistical modeling [19]. In [20], bandwidth extension based on classical signal processing has been applied to in-ear own voice signals for reconstructing high-frequency content. However, the quality of own voice processed by classical approaches is typically limited since body-conduction of own voice is difficult to account for. Many recent OVR approaches using body-conduction microphones or in-ear microphones are based on deep learning [13, 21–31]. Most deep learning-based approaches are trained using large amounts of device-specific recorded own voice signals, such as the ESMB corpus [26] or the VibraVox corpus [32]. However, since recording a

¹ Most approaches proposed and validated for body-conduction microphones can also be applied to in-ear microphones, as considered in this paper.

large amount of own voice signals requires a lot of recording effort, it has also been proposed to perform data augmentation by simulating own voice signals to enable training with less recording effort [13–15, 33, 34]. Although data augmentation by simulating microphone signals is quite common, e.g., to simulate different room characteristics [35–37], it should be realized that simulating bone conduction or in-ear own voice microphone signals is more challenging, since the own voice transfer characteristics are typically speech-dependent, device-specific and individual. In [33], it was proposed to simulate bone-conduction microphone own voice signals using a deep neural network (DNN) in an adversarial training paradigm. While the presented semi-supervised scheme enabled to reduce the amount of recorded own voice signals by half without sacrificing performance compared to supervised training with the full dataset, the performance deteriorated when the amount of recorded signals was reduced further. In [13], a speech-independent data augmentation technique was proposed to train an OVR system by simulating in-ear own voice signals based on relative transfer functions (RTFs) between an outer microphone and an in-ear microphone. The performance increased by introducing variance to the simulated own voice signals by considering RTFs estimated from different segments and different talkers. Additionally, the performance notably improved by fine-tuning with recorded own voice signals after training with simulated own voice signals. In [15], a similar data augmentation technique as in [13] was used to simulate bone-conduction signals. Instead of using RTFs from different segments and talkers, in [14] it was proposed to introduce additional variance to the simulated own voice signals by adding random values to the magnitude of the RTF estimated from a single talker during training, resulting in a performance increase. In [34], it was proposed to perform talker-specific fine-tuning after training with simulated data, obtained with speech-independent data augmentation techniques similar to [13–15]. A considerable performance gain from fine-tuning was observed even when only few recorded own voice signals were available.

While previous studies in [13–15, 34] have investigated the use of speech-independent data augmentation techniques, in this paper we investigate speech-dependent data augmentation techniques to train an OVR system using simulated in-ear own voice signals. The proposed data augmentation techniques (see Fig. 3.1) use a small amount of own voice signals recorded with several talkers wearing a hearable device [38] to first estimate models of own voice transfer characteristics [16]. These models can then be used to simulate a large amount of in-ear own voice signals from a dataset of clean speech signals to train an OVR system. In this paper, we will compare data augmentation using (individual or talker-averaged) speech-dependent models with speech-independent models. We will consider three different training procedures: training with a small amount of recorded signals, training with a large amount of simulated signals, and fine-tuning with recorded signals after training with simulated signals. For the three considered training procedures, we investigate the influence of the amount of recorded own voice signals, both in terms of number of talkers as well as number of utterances per talker on the OVR performance. Experiments are carried out for an OVR system based on the joint spatial and tempo-spectral non-linear filter (FT-JNF) architecture [39], using both the outer and in-ear microphone signal as input to estimate clean broadband speech. Results

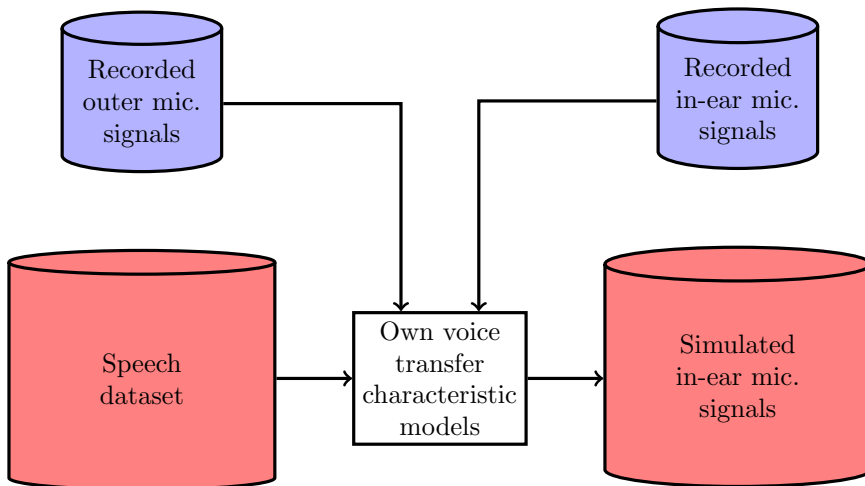


Figure 3.1: Data augmentation with own voice transfer characteristic models, which are estimated from recorded outer and in-ear microphone signals. During data augmentation, in-ear own voice signals are simulated from a speech dataset using the estimated own voice transfer characteristic models.

show that the proposed speech-dependent data augmentation results in higher OVR performance than speech-independent data augmentation. Speech-dependent individual data augmentation outperforms training using only the recorded own voice signals, while performance can still be considerably improved by additional fine-tuning. In addition, results show that the number of recorded utterances per talker has a smaller influence on the OVR performance than the number of recorded talkers. Moreover, the results demonstrate that the proposed speech-dependent individual data augmentation is still highly effective even when only a limited amount of recorded own voice signals is available.

The remainder of this paper is organized as follows. In Section 3.2, the signal model is introduced and the considered speech-independent and speech-dependent own voice transfer characteristic models are presented. Based on these models, in Section 3.3 several data augmentation techniques are proposed to simulate in-ear own voice signals to train an OVR system. In Section 3.4, the experimental evaluation setup for investigating the proposed data augmentation techniques and training procedures is described. In Section 3.5, the experimental results are presented and discussed.

3.2 Own voice transfer characteristic models

After introducing the signal model in Section 3.2.1, speech-independent and speech-dependent own voice transfer characteristic models are presented in Section 3.2.2.

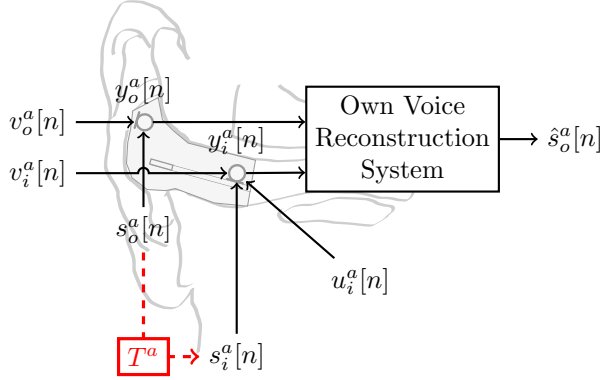


Figure 3.2: Block diagram of a multi-microphone OVR system using an outer and an in-ear microphone of a hearable device. The own voice transfer characteristics of talker a between the outer and in-ear microphone are represented by T^a .

3.2.1 Signal model

Consider a hearable device equipped with an outer microphone and an in-ear microphone, as depicted in Fig. 3.2. The microphone signals are denoted by subscripts o for the outer microphone and i for the in-ear microphone, respectively. We assume that the hearable is worn by a person (referred to as talker) in a noisy environment. The time-domain signals $s_o^a[n]$ and $s_i^a[n]$ denote the own voice of talker a at the outer microphone and the in-ear microphone, respectively, where n denotes the discrete-time index. The outer microphone signal $y_o^a[n]$ consists of the own voice component $s_o^a[n]$ and the environmental noise component $v_o^a[n]$ (including sensor noise), i.e.,

$$y_o^a[n] = s_o^a[n] + v_o^a[n]. \quad (3.1)$$

Similarly, the in-ear microphone signal $y_i^a[n]$ consists of the own voice component $s_i^a[n]$, the environmental noise component $v_i^a[n]$ (including sensor noise), and body-produced noise $u_i^a[n]$ (e.g., respiratory and heart sounds), i.e.,

$$y_i^a[n] = s_i^a[n] + v_i^a[n] + u_i^a[n]. \quad (3.2)$$

In practice, the recorded own voice signal at the in-ear microphone also contains body-produced noise. The relationship between the own voice components of talker a at the outer microphone and the in-ear microphone is referred to as the own voice transfer characteristics $T^a\{\cdot\}$, i.e.,

$$s_i^a[n] = T^a\{s_o^a[n]\}. \quad (3.3)$$

3.2.2 Own voice transfer characteristic models

In [13–16], several models for the own voice transfer characteristics $T^a\{\cdot\}$ have been proposed. The transfer characteristics are either modeled as a time-invariant or a time-varying linear system, either for each individual talker or averaged over all talkers. In the short-time Fourier transform (STFT) domain, the time-varying system is approximated² as

$$S_i^a(k, l) = H^a(k, l) \cdot S_o^a(k, l), \quad (3.4)$$

where k denotes the frequency bin index, l denotes the time frame index and $H^a(k, l)$ denotes the time-varying RTF of talker a between the outer microphone and the in-ear microphone. Assuming that own voice recordings of multiple talkers in a noiseless environment are available and sensor noise can be neglected compared to the own voice components in both microphone signals, i.e., $v_o^a[n] \approx 0$ and $v_i^a[n] \approx 0$, the following speech-independent and speech-dependent models have been proposed:

1. *Speech-independent models:* If the own voice transfer characteristics are assumed to be independent of the phonetic content, the transfer characteristics of talker a can be modeled as a time-invariant RTF $H^a(k)$. Considering all active STFT frames of the recorded microphone signals of talker a , the least-squares RTF estimate [40] is given by

$$\hat{H}^a(k) = \frac{\sum_l Y_i^a(k, l) \cdot Y_o^{a,*}(k, l)}{\sum_l |Y_o^a(k, l)|^2}, \quad (3.5)$$

where $\cdot^*$ denotes complex conjugation and it is assumed that the own voice component at the outer microphone and the body-produced noise at the in-ear microphone are independent. Since this speech-independent model is based only on recorded microphone signals of talker a , it is an individual model. A speech-independent talker-averaged model can be obtained by performing RTF estimation using all STFT frames of the recorded microphone signals of all talkers, i.e.,

$$\hat{H}^{\text{avg}}(k) = \frac{\sum_a \sum_l Y_i^a(k, l) \cdot Y_o^{a,*}(k, l)}{\sum_a \sum_l |Y_o^a(k, l)|^2}. \quad (3.6)$$

2. *Speech-dependent models:* Since the own voice transfer characteristics in general depend on the phonetic content, it has been proposed in [16] to model the transfer characteristics using a time-varying model, assuming that the transfer characteristics for each phoneme can be modeled as a time-invariant RTF. First, the frame-wise phoneme sequence $p_o^a(l) \in \{1, \dots, P\}$, with P possible phoneme classes, is obtained from the (noiseless) outer microphone

²This approximation is only time-varying between STFT frames and not within a single STFT frame. Circular convolution effects are also neglected in this approximation, but can be reduced by appropriate windowing.

signal $y_o^a[n]$ using a phoneme recognition system. The RTF for phoneme p' is then estimated from all STFT frames in which this phoneme is present as

$$\hat{H}_{p'}^a(k) = \frac{\sum_{p_o^a(l)=p'} Y_i^a(k, l) \cdot Y_o^{a,*}(k, l)}{\sum_{p_o^a(l)=p'} |Y_o^a(k, l)|^2}, \quad p' = 1, \dots, P. \quad (3.7)$$

Since this model is based only on recorded microphone signals of talker a , it is a speech-dependent individual model. A speech-dependent talker-averaged model can be obtained by performing RTF estimation using all STFT frames of the recorded microphone signals of all talkers in which phoneme p' is present, i.e.,

$$\hat{H}_{p'}^{\text{avg}}(k) = \frac{\sum_a \sum_{p_o^a(l)=p'} Y_i^a(k, l) \cdot Y_o^{a,*}(k, l)}{\sum_a \sum_{p_o^a(l)=p'} |Y_o^a(k, l)|^2}, \quad p' = 1, \dots, P. \quad (3.8)$$

Although this speech-dependent talker-averaged model does not account for individual differences, it has been shown in [16] that on average it is able to estimate in-ear own voice signals of different talkers (i.e., under talker mismatch) better than speech-dependent individual models.

3.3 Data augmentation techniques using own voice transfer characteristic models

In order to train a deep learning-based OVR system aiming at estimating clean broadband speech from both the (noisy) outer microphone signal and the (band-limited) in-ear microphone signal, a large amount of own voice signals is required. Instead of recording a large amount of own voice signals, in this section we propose several data augmentation techniques using the own voice transfer characteristic models presented in Section 3.2.2. These models can be used to simulate in-ear own voice signals and can be estimated from a relatively small amount of recorded own voice signals (see Fig. 3.1). In addition to speech-independent data augmentation, we propose two phoneme-dependent data augmentation techniques to simulate in-ear own voice signals, either using matching phoneme sequences or random phoneme sequences. While several speech-independent data augmentation techniques have been proposed in the literature [13–15], our goal is not to benchmark all of them. Instead, we have considered a representative speech-independent baseline that conceptually aligns with [13], where speech-independent RTFs are used to generate training data. An overview of the proposed data augmentation techniques is presented in Table 3.1, also indicating the time-variance and the number of used RTFs. For all data augmentation techniques, the in-ear own voice signals are simulated from large datasets of clean speech signals (e.g., LibriSpeech [41], Common Voice [42]) instead of from recorded outer microphone own voice signals. This has the advantage that simulating a large amount of in-ear own voice signals is easily feasible, even though no individual transfer characteristic models are available for the talkers in these datasets. To obtain an augmented in-ear own voice signal $\hat{s}_i[n]$, a speech signal $s[n]$ from the dataset is selected as the outer microphone own

Table 3.1: Overview of the proposed augmentation techniques.

Augmentation technique		Time-variance	Number of RTFs
Speech-independent	Individual	Time-invariant	Number of talkers
	Talker-averaged	Time-invariant	1
Speech-dependent	Individual	Phoneme-dependent	$P \times \text{Number of talkers}$
	Talker-averaged	Phoneme-dependent	P
Random phoneme	Individual	Random	$P \times \text{Number of talkers}$

voice signal. In the STFT domain, the augmented in-ear own voice signal $\hat{S}_i(k, l)$ is generated based on (3.4), i.e.,

$$\hat{S}_i(k, l) = \hat{H}(k, l) \cdot S(k, l), \quad (3.9)$$

where depending on the data augmentation technique, different RTF estimates $\hat{H}(k, l)$ are used (see below). Finally, the augmented in-ear own voice signal $\hat{s}_i[n]$ is computed by transforming $\hat{S}_i(k, l)$ back to the time domain using a weighted overlap-add (WOLA) scheme. We will now discuss the considered data augmentation techniques in more detail:

1. *Speech-independent augmentation:* When performing data augmentation using speech-independent individual models, the augmented in-ear own voice signals are generated using time-invariant RTFs as

$$\hat{S}_i(k, l) = \hat{H}^a(k) \cdot S(k, l). \quad (3.10)$$

For each speech signal from the dataset, the RTF estimate $\hat{H}^a(k)$ in (3.5) of a random talker a is chosen (for the entire signal) from all available talkers, similarly as in [13]. When performing data augmentation using the speech-independent talker-averaged model, the RTF estimate $\hat{H}^{\text{avg}}(k)$ in (3.6) is used instead of $\hat{H}^a(k)$ in (3.10), i.e., the same time-invariant RTF is used for all augmented in-ear own voice signals.

2. *Speech-dependent augmentation:* When performing data augmentation using speech-dependent individual models, the frame-wise phoneme sequence $p(l)$ is first determined for the speech signal $s[n]$ and then used to select the matching phoneme-specific RTF estimate $\hat{H}_{p(l)}^a(k)$ in (3.7) for each STFT frame. Similarly as for speech-independent individual augmentation, a random talker a is chosen for the entire signal. To prevent discontinuities during phoneme transitions, recursive smoothing is applied to the RTF estimates, i.e.,

$$\tilde{H}_{p(l)}^a(k) = \alpha \cdot \tilde{H}_{p(l-1)}^a(k) + (1 - \alpha) \cdot \hat{H}_{p(l)}^a(k), \quad (3.11)$$

where $\tilde{H}_{p(l)}^a(k)$ denotes the smoothed RTF of talker a for frame l , and α denotes the (time- and frequency-independent) smoothing constant. Using the smoothed RTFs, the augmented in-ear own voice signals are generated as

$$\hat{S}_i(k, l) = \tilde{H}_{p(l)}^a(k) \cdot S_o(k, l). \quad (3.12)$$

When performing data augmentation using the speech-dependent talker-averaged model, the talker-averaged phoneme-specific RTFs $\tilde{H}_{p(l)}^{\text{avg}}(k)$ in (3.8) are used instead of $\hat{H}_{p(l)}^a(k)$ in (3.11). Compared to speech-independent augmentation, speech-dependent augmentation does not only introduce time-varying speech-dependent behavior, but also incorporates additional variance into the training dataset by using P times more phoneme-specific RTFs (see Table 3.1).

3. *Random phoneme individual augmentation:* To investigate the influence of speech-dependency and additional variance in the training dataset separately, we also consider an individual augmentation technique using random RTF selection instead of speech-dependent RTF selection, i.e., instead of obtaining the phoneme sequence $p(l)$ from $s[n]$, the phoneme sequence is randomly generated. This augmentation technique uses the same amount of RTFs as speech-dependent individual augmentation, but applies the RTFs of random phonemes instead of matching phonemes³.

In the experimental evaluation (see Sections 3.4 and 3.5), we will compare the performance of the presented data augmentation techniques when training an OVR system using a large amount of simulated own voice signals. In addition, we will investigate the influence of the number of recorded talkers and recorded utterances per talker, which are used to estimate the own voice transfer characteristic models required for data augmentation.

3.4 Experimental setup

This section describes the experimental setup used to evaluate the influence of the proposed data augmentation techniques on the performance of an OVR system. Section 3.4.1 provides details on the speech and noise datasets used for training and evaluation. Section 3.4.2 describes the DNN architecture of the OVR system. Section 3.4.3 describes the training procedures and the hyperparameters, and Section 3.4.4 discusses the experimental evaluation conditions and performance metrics.

3.4.1 Datasets

This section discusses the speech and noise datasets used for estimating the own voice transfer characteristic models and for training and evaluating the OVR sys-

³Although it would also be possible to consider random phoneme talker-averaged augmentation, this would incorporate less variance into the training dataset than random phoneme individual augmentation, such that we chose not to investigate it.

Table 3.2: Overview of the speech and noise datasets used in the experimental evaluation. The speech and the noise datasets are disjunct between training/validation/fine-tuning and evaluation. All dataset splits marked by ¹ and all splits marked by ² are identical, respectively.

Use of dataset	Speech (own voice)	Noise
Estimation of transfer characteristic models	Recorded own voice signals (training/validation set, max. 12/2 talkers, max. 306 utterances, max. 6/1 hours) ¹	-
Training/validation (recorded signals)	Recorded own voice signals (training/validation set, max. 12/2 talkers, max. 306 utterances, max. 6/1 hours) ¹	Individually spatialized noise (120.7/20.1 hours) ²
Training/validation (augmented signals)	Augmented own voice signals (115.7 hours)	Individually spatialized noise (120.7/20.1 hours) ²
Fine-tuning	Recorded own voice signals (training/validation set, max. 12/2 talkers, max. 306 utterances, max. 6/1 hours) ¹	Individually spatialized noise (120.7/20.1 hours) ²
Evaluation	Recorded own voice signals (test set, 4 talkers, 306 utterances, max. 2 hours)	Individually spatialized noise (40.2 hours)

tem. The recordings and the experimental validation were conducted at a sampling frequency of 16 kHz. An overview of the datasets and their usage is presented in Table 3.2.

3.4.1.1 Recorded own voice signals

Own voice signals from 18 native German talkers (5 female, 13 male) were recorded (see [16] for details). The talkers were wearing the closed-vent variant of the Hearpiece [38], a prototype hearable device with an in-ear microphone and multiple outer microphones, of which the *Concha* microphone was chosen as the outer microphone in the experiments. Both microphones are Knowles SPH1642HT5H-1 MEMS microphones. For each talker, utterances of 306 phonetically balanced and representative German sentences were recorded, corresponding to approximately 25 to 30 minutes of own voice recordings per talker (approximately 9 hours in total). Although own voice recordings were performed at both ears, only the recordings at the left ear were considered here. The recorded own voice signals were split into disjoint training, validation and test sets of 12, 2, and 4 talkers, respectively. Training and validation sets were used for estimating transfer characteristic models for data aug-

mentation and for training directly on the recorded signals. The test set was used exclusively for evaluating the OVR systems.

To estimate the own voice transfer characteristic models from the recorded own voice signals, we followed the procedure described in Section 3.2.2. RTFs were estimated using either all utterances available for each talker (individual) or all utterances available for all talkers (talker-averaged) in the training set. For the speech-dependent models, a proprietary in-house phoneme recognition system was used, which was trained with about 1000 hours of German speech and distinguishes between $P = 62$ phoneme classes. The system is similar to the system used in [43] and uses Mel-Filterbank features as input to a temporal convolutional neural network architecture. The system provides phoneme labels with 30 ms temporal resolution, which are matched to STFT frames based on minimal temporal distance.

Since the recorded in-ear own voice signals are band-limited, the RTFs were estimated at a sampling frequency of 5 kHz, by first downsampling the microphone signals. The RTFs were estimated in an STFT framework with a frame length of 25.6 ms (128 samples), 50 % overlap, and a square-root Hann analysis window. After simulation, the augmented in-ear own voice signals were upsampled to 16 kHz again. The same procedure is carried out separately for the validation set to estimate own voice transfer characteristic models used in validation.

3.4.1.2 *Augmented own voice signals*

The estimated own voice transfer characteristic models were used for data augmentation to obtain simulated in-ear own voice signals (see Fig. 3.1). Since own voice data augmentation is only carried out for training and validation, RTFs from the previous subsection are estimated using only the training and validation subsets, amounting to approx. 6/1 hours, respectively (see Table 3.2). For the outer microphone signals, clean speech signals from the Common Voice dataset [42] were used (German part, v11.0, only *validated* subset). While the Common Voice dataset consists of 1157 hours of speech signals, only 10 % of this dataset (corresponding to 115.7 hours) was used to construct the augmented own voice dataset, since preliminary experiments suggest that this amount is sufficient for training an OVR system. Several data augmentation techniques were considered (see Section 3.3), namely speech-independent augmentation (individual and talker-averaged), speech-dependent augmentation (individual and talker-averaged) and random phoneme augmentation (individual). The models were applied at a sampling frequency of 5 kHz, by downsampling the clean speech signals, and using the same STFT framework as for model estimation. For speech-dependent augmentation, the same phoneme recognition system as for model estimation was used to determine the frame-wise phoneme sequence from the clean speech signals. Smoothing of the RTFs in (3.11) was carried out with $\alpha = 0.8$ for both speech-dependent and random phoneme augmentation. If the speech-dependent and random phoneme augmentation techniques encountered a phoneme during simulation for which no RTF estimate is available, a fallback RTF averaged over all available phoneme RTF estimates was used. This may happen when only few recorded utterances are available to estimate the own voice transfer characteristics (as in Section 3.5.3). The augmented in-ear own voice sig-

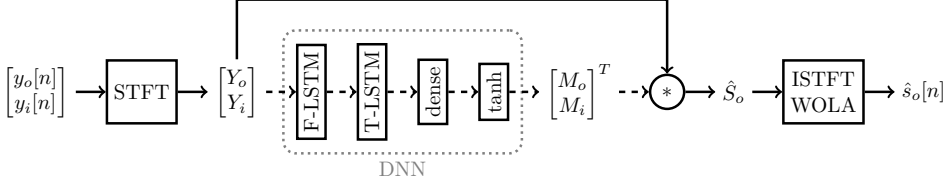


Figure 3.3: Architecture of the OVR system.

nals in the STFT domain were transformed back to the time domain using a WOLA scheme with a square-root Hann synthesis window and were upsampled to 16 kHz. The same procedure is carried out separately with the own voice transfer characteristic models for the validation set to obtain simulated in-ear own voice signals for validation.

3.4.1.3 Environmental noise

The environmental noise used for training, validation, and testing is a spatialized version of the noise dataset from the fifth DNS challenge [44], consisting of approximately 181 hours of environmental noise. Since it was shown in [28] that training an OVR system with individually spatialized noise signals generalizes well to recorded noise signals, individually measured device-specific transfer functions from different directions were used to generate the noise components in the outer and in-ear microphones for each talker. The transfer functions between 8 loudspeakers and both microphones were measured for all talkers using exponential sweeps. The loudspeakers were positioned at approximately 1.5 m distance from the talker in 45°-steps in the horizontal plane.

Half of the spatialized noise dataset consisted of point noise sources, while the other half consisted of pseudo-diffuse noise. To generate the point noise source signals, a random noise sample from the DNS dataset was filtered with the device-specific transfer functions corresponding to a random direction. To generate the pseudo-diffuse noise signals, time-shifted copies of a random noise sample from the DNS dataset were filtered with the transfer functions for the all directions, and then added. To account for body-produced noise, which is present in the recorded own voice signals, white noise was added to the spatialized in-ear signal with a random level uniformly distributed in the range of $[-\infty, -60]$ dB relative to the in-ear environmental noise signal.

The noise dataset was split into equal parts (approximately 10.1 hours per talker), so that different noise types are spatialized with individually measured impulse responses for each talker. This way, different noise types are included in training and validation than in evaluation.

Table 3.3: Number of parameters and MACs per second for each layer in the DNN architecture used in the OVR system. M indicates millions, G indicates billions.

Layer	Parameters	MACs/s
F-LSTM	1.06 M	17.14 G
T-LSTM	0.33 M	5.31 G
Dense	516	32 508
Total	1.39 M	22.45 G

3.4.2 Own voice reconstruction system

Fig. 3.3 depicts the considered OVR system, which is based on the FT-JNF DNN architecture [39]. The noisy microphone signals $y_o[n]$ and $y_i[n]$ are transformed into the STFT domain with a frame length of 32 ms (512 samples), 50 % overlap and a square-root Hann analysis window. The (broadband) own voice component in the outer microphone signal is then estimated by applying complex-valued masks to the outer and in-ear microphone signals, i.e.,

$$\hat{S}_o(k, l) = M_o(k, l) \cdot Y_o(k, l) + M_i(k, l) \cdot Y_i(k, l), \quad (3.13)$$

where $M_o(k, l)$ and $M_i(k, l)$ denote the mask for the outer and in-ear microphone, respectively. Even though the in-ear microphone signal only contains band-limited own voice, it was shown in [28] that combining both microphone signals leads to a higher performance than only applying a mask to the outer microphone signal due to a better noise reduction in the lower frequencies. The inputs to the DNN are the real and imaginary parts of the complex-valued STFT coefficients $Y_o(k, l)$ and $Y_i(k, l)$. The DNN architecture consists of a uni-directional long short-term memory (LSTM) layer with 512 hidden units operating along the frequency dimension (F-LSTM), followed by a uni-directional LSTM layer with 128 hidden units operating along the time dimension (T-LSTM). This is followed by a dense layer combining the 128 outputs of the second LSTM layer to 4 outputs, followed by a tanh activation. The outputs of the DNN are the real and imaginary parts of the complex-valued masks $M_o(k, l)$ and $M_i(k, l)$. Overall, the OVR system consists of approximately 1.39 million parameters, requires about 22.45 billion multiply-accumulate operations (MACs) per second⁴. Table 3.3 provides an overview of the complexity of the DNN required for each individual layer. The estimated STFT coefficient in (3.13) is transformed back to the time-domain signal $\hat{s}_o[n]$ using a WOLA scheme with a square-root Hann synthesis window.

⁴Related work in [45] investigates low-complexity versions of the OVR system using speech-dependent data augmentation proposed in this paper. Results showed that good performance could still be obtained by a system with only 31 k parameters and a complexity of 0.5 G MACs per second.

3.4.3 Training procedures

The OVR system was trained on mixtures of recorded or augmented own voice signals and spatialized environmental noise signals, truncated to a signal length of 3 s. The own voice and noise components $S_o(k, l)$ and $V_o(k, l)$ in the outer microphone signal were added as

$$Y_o(k, l) = S_o(k, l) + q \cdot V_o(k, l), \quad (3.14)$$

where the scaling factor q determines the signal-to-noise ratio (SNR). The noise component in the in-ear microphone $V_i(k, l)$ was scaled by the same factor q as $V_o(k, l)$, i.e.,

$$Y_i(k, l) = (S_i(k, l) + U_i(k, l)) + q \cdot V_i(k, l), \quad (3.15)$$

where $U_i(k, l)$ denotes body-produced noise present during own voice recording. Due to device attenuation, the noise component in the in-ear microphone signal $V_i(k, l)$ has a much lower level than the noise component in the outer microphone signal $V_o(k, l)$, leading to a higher SNR at the in-ear microphone than at the outer microphone. The procedure in (3.14)-(3.15) maintains the SNR difference between both microphones. During training, uniformly distributed SNRs between -10 dB and 25 dB were considered at the outer microphone.

Own voice signals and environmental noise signals of the same talker were added, so that they were individually matched. As loss function, the combined L_1 loss in time domain and STFT domain (after re-analysis) [46] between the STFT magnitudes was used, i.e.,

$$L = \sum_n |s_o[n] - \hat{s}_o[n]| + \sum_{k,l} \left| |\text{STFT}\{s_o[n]\}(k, l)| - |\text{STFT}\{\hat{s}_o[n]\}(k, l)| \right|, \quad (3.16)$$

where $s_o[n]$ and $\hat{s}_o[n]$ denote the clean and estimated own voice signal at the outer microphone, respectively. The same STFT parameters as in DNN processing were used to compute the STFT term in the loss. During training, mean-variance normalization was applied to the noisy microphone signals independently for each microphone. As in [47], the mean and variance of the noisy outer microphone signals was also used to scale the corresponding clean own voice components, which serve as the training targets. Each training batch consisted of four utterances. In addition to training with augmented own voice signals, we also investigate additional fine-tuning with recorded own voice signals afterwards. Here, we refer to fine-tuning as further training (selected layers of) a previously trained DNN with a smaller learning rate. We investigate fine-tuning of the dense layer, the T-LSTM layer, the F-LSTM layer, or all layers of the FT-JNF architecture.

For all training procedures, we used the ADAM optimizer [48]; for training with recorded and augmented own voice signals an initial learning rate of 10^{-4} was used, whereas for fine-tuning a smaller initial learning rate of 10^{-5} was used. Training proceeded up to a maximum of 100 epochs, where one epoch corresponds to iterating over the own voice training set once. The learning rate was halved after three consecutive epochs without validation loss improvement, and early stopping was applied after six consecutive epochs without validation loss improvement. The

PyTorch [49] (v1.10) framework was used, and computations were executed on two NVIDIA GeForce RTX 2080 SUPER GPUs.

While for most training conditions the recorded own voice signals of all 12 talkers with 306 utterances each were used in training, in Section 3.5.3 the influence of the number of talkers and the number of recorded utterances per talker is investigated. In these experiments, the numbers of recorded talkers considered are equal to 1, 2, 3, 4, 6, 8, 10, and 12 (keeping the number of utterances fixed to 306), while the numbers of recorded utterances per talker considered are equal to 1, 3, 6, 12, 25, 75, 150, and 306 (keeping the number of talkers fixed to 12).

3.4.4 *Evaluation details*

The performance of the OVR systems trained using different procedures was evaluated on mixtures of recorded own voice signals and spatialized environmental noise signals. As mentioned before, the test set consists of recorded own voice signals (306 utterances) from 4 talkers, which are different from the talkers during training. The performance was evaluated at SNRs of -10, -5, 0, 5, and 10 dB in the outer microphone signal. As in training, own voice signals and environmental noise signals of the same talker are added, so that they are individually matched. As evaluation metrics, we considered the wide-band perceptual evaluation of speech quality (PESQ) [50] and short-time objective intelligibility (STOI) [51] metrics. These metrics were computed for the estimated own voice signal, using the clean own voice component in the outer microphone signals as the reference.

As baseline OVR algorithm, we consider the extreme bandwidth extension network (EBEN) [14], trained with augmented signals and additional fine-tuning with recorded signals. Different from [14], we only train the generator network in a fully supervised scheme with the same loss function in (3.16). It should be noted that EBEN only uses the in-ear microphone signal as input.

3.5 Experimental results

In this section, the results of the experimental evaluation are presented and discussed. Section 3.5.1 compares the influence of the data augmentation techniques on the OVR performance when training with a large amount of augmented data. The results demonstrate that the proposed speech-dependent individual augmentation yields the best performance, outperforming a system trained using only recorded own voice signals. Section 3.5.2 investigates the potential performance improvement from fine-tuning different layers of the DNN trained using speech-dependent individual augmentation. Section 3.5.3 explores the trade-off between recording effort and performance for systems trained using recorded own voice signals and systems trained using augmented own voice signals (with or without fine-tuning), both in terms of number of recorded talkers and number of recorded utterances per talker. Audio examples supporting the experimental results are available online⁵.

⁵Audio examples [online]: https://m-ohlenbusch.github.io/own_voice_augmentation_examples/

3.5.1 *Influence of data augmentation techniques*

For different SNRs, Fig. 3.4 shows the performance in terms of PESQ and STOI for systems trained using speech-independent and speech-dependent data augmentation (individual and talker-averaged), as well as a system trained using random phoneme individual data augmentation and a system trained using only recorded own voice signals. The performance metrics for the unprocessed outer and in-ear microphone signals are also shown. Circles indicate average results and error bars indicate standard deviations over the test set examples. It should be noted that all OVR systems considered in this section use the entire training and validation dataset of recorded own voice signals (amounting to approx. 6/1 hours), either directly or to generate 115.7 hours of augmented data (see Table 3.2).

It can be observed that the proposed speech-dependent augmentation consistently leads to higher performance than speech-independent augmentation for both individual as well as talker-averaged augmentation, with speech-dependent individual augmentation yielding the highest performance in terms of both metrics. In addition, random phoneme individual augmentation leads to higher performance than speech-independent individual augmentation. This can be explained by the additional variance from augmenting with multiple RTFs per talker and matches previously reported results in [13, 14], where an increase in the number of RTFs or introducing additional variance in augmentation improved performance. When comparing random phoneme individual augmentation with speech-dependent individual augmentation, it can be observed that using phoneme-matched RTF selection instead of random RTF selection leads to better results in terms of PESQ and similar results in terms of STOI. This indicates that speech-dependent modeling yields a benefit in addition to the performance gained by augmenting with multiple RTFs per talker. Moreover, the results show that both random phoneme as well as speech-dependent individual data augmentation outperform a system trained on recorded own voice signals (except in terms of STOI at high SNR).

In summary, training an OVR system using speech-dependent individual augmentation technique to simulate a large amount of own voice signals yields the best performance among all considered data augmentation techniques. In the following sections, we will therefore only consider speech-dependent individual augmentation as data augmentation technique.

3.5.2 *Influence of fine-tuning*

This section investigates the potential performance improvement by fine-tuning different layers of the DNN architecture (see Fig. 3.3) with recorded own voice signals after training the OVR system using speech-dependent individual augmentation. Similarly as in the previous section, the entire training and validation dataset of recorded own voice signals (amounting to approx. 6/1 hours) is used to generate 115.7 hours of augmented data, as well as for fine-tuning. For different SNRs, Fig. 3.5 shows the PESQ and STOI improvement compared to the noisy outer microphone signal obtained by training only with recorded own voice signals, training only with

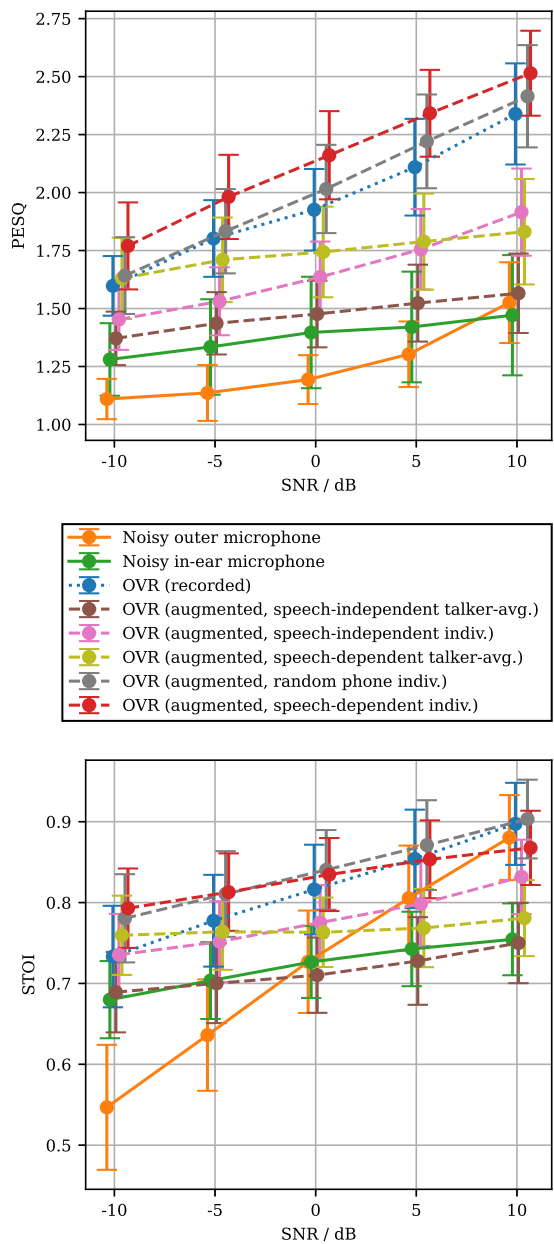


Figure 3.4: OVR performance in terms of PESQ (top) and STOI (bottom) for different SNRs, achieved by systems trained either only with recorded own voice signals directly or with simulated own voice signals, obtained through different data augmentation techniques. Data points are slightly shifted horizontally to improve readability.

augmented own voice signals (without fine-tuning), or performing additional fine-tuning of different layers (dense layer, T-LSTM layer, F-LSTM layer, all layers). All fine-tuning approaches lead to higher or equal performance than training only with augmented own voice signals, and significantly better results than training only with recorded signals. While hardly any improvement is obtained by fine-tuning only the dense layer, larger improvements are obtained by fine-tuning only the T-LSTM layer or the F-LSTM layer, and the largest improvements are obtained by fine-tuning all layers.

In summary, the results show that fine-tuning with recorded own voice signals is beneficial in addition to augmented training using transfer characteristic models obtained from the recorded own voice signals. In the following section, only fine-tuning of the full DNN is further considered.

3.5.3 *Influence of amount of recorded own voice signals*

In Sections 3.5.1 and 3.5.2 the entire training dataset of recorded own voice signals (i.e., 12 talkers, 306 utterances, see Section 3.4.1) was used for estimating own voice transfer characteristic models as well as for fine-tuning. In this section, the trade-off between OVR performance and the amount of recorded own voice signals used is investigated, both in terms of number of talkers and number of utterances per talker. Three training conditions are compared: training only with recorded own voice signals, training only with augmented own voice signals using speech-dependent individual augmentation, and performing additional fine-tuning (all layers) with recorded own voice signals. It should be noted that irrespective of the assumed amount of recorded own voice signals, 115.7 hours of augmented own voice signals are generated. In addition, we also compare to EBEN as a baseline algorithm.

For the considered training conditions, Fig. 3.6 shows the PESQ and STOI improvement compared to the noisy outer microphone signal for different number of talkers (306 utterances per talker) and different number of utterances per talker (12 talkers). Circles indicate average results and error bars indicate standard deviations over the test set examples and SNRs. For all considered number of talkers and number of utterances per talker, fine-tuning with recorded own voice signals clearly leads to higher performance compared to training only with augmented own voice signals and training only with recorded own voice signals. Training only with augmented own voice signals leads to higher performance than training only with recorded own voice signals (except for Δ PESQ with number of talkers = 1). Moreover, all proposed OVR systems using both microphone signals as input clearly outperform the EBEN baseline algorithm using only the in-ear microphone signal as input.

Investigating the influence of the number of recorded talkers, it can be observed that the performance for all three training conditions generally decreases when the number of talkers is reduced. When training only with augmented own voice signals, the PESQ scores drop when there are fewer than 4 talkers. When performing additional fine-tuning, the PESQ scores already drop when there are fewer than 8 talkers. While the improvements achieved by fine-tuning over only augmented training are larger for more talkers, the improvements achieved by augmented training

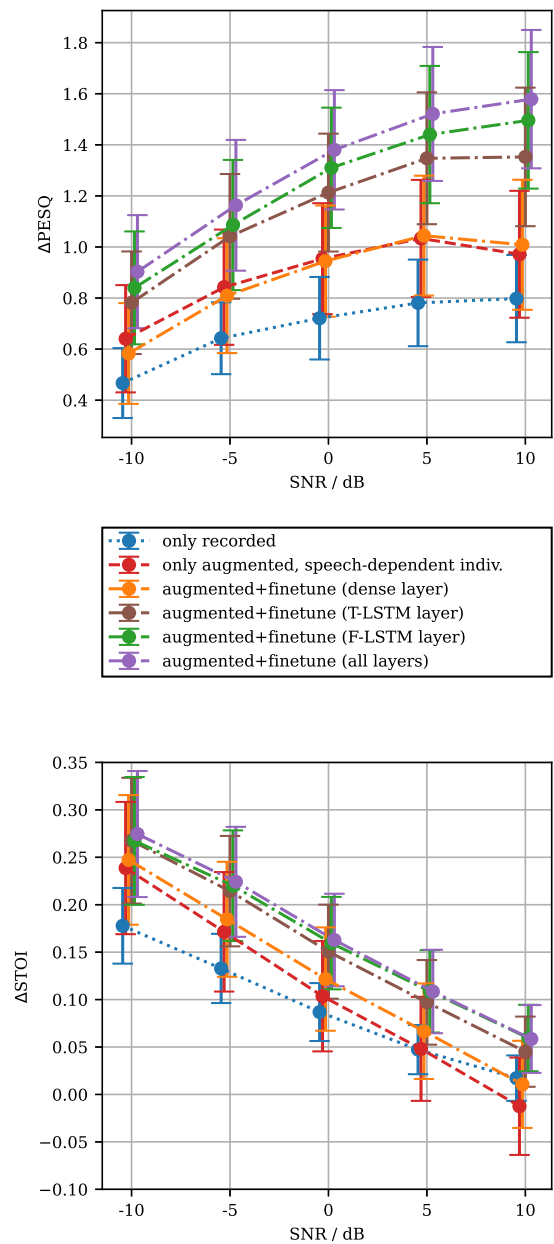


Figure 3.5: PESQ (top) and STOI (bottom) improvement for different SNRs, achieved by the OVR system trained only with recorded own voice signals, trained only with augmented own voice signals, or by performing additional fine-tuning of different layers with recorded own voice signals. Data points are slightly shifted horizontally to improve readability.

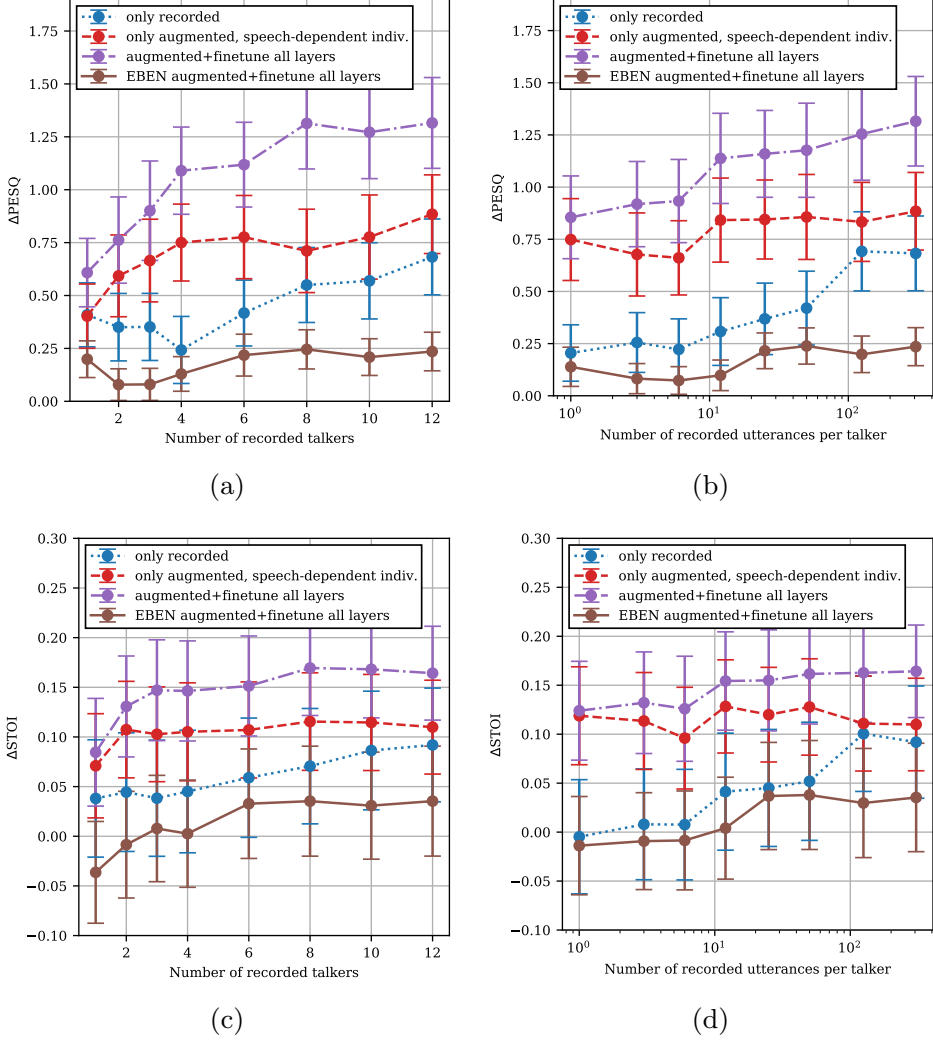


Figure 3.6: PESQ and STOI improvement achieved by the OVR system trained with a different number of recorded talkers (a), (c), and a different number of recorded utterances per talker (b), (d). Three training conditions are considered: training only with recorded own voice signals, training only with augmented own voice signals, and performing additional fine-tuning with recorded own voice signals. The performance of EBEN is also included for comparison. Results were averaged over SNRs of -10, -5, 0, 5, and 10 dB.

over training with recorded own voice signals are larger for fewer talkers (except for a single talker). These results suggest that the proposed speech-dependent individual data augmentation technique is particularly effective when recorded own voice signals of few talkers are available, and additional fine-tuning is more effective when recorded own voice signals of more talkers are available.

Investigating the influence of the number of utterances per talker, it can be observed that the performance for all three training conditions generally decreases when the number of utterances is reduced. When training with recorded own voice signals, the performance strongly decreases, whereas when training with augmented own voice signals, the performance only slightly decreases. When fine-tuning with recorded own voice signals, there is only a small benefit over only augmented training when few utterances per talker are available, but a larger benefit when more utterances per talker are available. The results show that augmented training achieves the largest benefit compared to training with recorded own voice signals when few utterances per talker are available. Since the performance for few utterances is still relatively high, it appears that recording more utterances per talker is less beneficial than recording more talkers for the proposed augmentation technique.

In summary, the results show that the proposed speech-dependent individual augmentation technique enables training an OVR system even when few recorded own voice signals are available. Increasing the recording effort can substantially improve performance, especially by increasing the number of recorded talkers.

3.6 Discussion

The experiments presented in Section 3.5 evaluated speech-dependent own voice data augmentation, fine-tuning with recorded signals, and the impact of the amount of recorded signals used for training an own voice reconstruction system.

The results demonstrate significant performance gains when using speech-dependent compared to speech-independent augmentation. A comparison with the random phoneme augmentation technique reveals that about two-thirds of the PESQ improvement can be attributed to introducing random variance to the training data, while the remaining improvement is due to modeling of speech-dependent behavior. In terms of STOI, the relative benefit of speech-dependent modeling is smaller, which may be due to differences in how PESQ and STOI measure the difference between processed signals and clean reference signals. Additionally, comparing speech-dependent individual augmentation to speech-dependent talker-averaged augmentation shows a clear advantage for individual modeling. Interestingly, this contrasts with the finding from [16], where talker-averaged models performed better in predicting in-ear signals of unseen talkers. This suggests that, in the context of data augmentation for training an OVR system, introducing variance by considering 12 individual RTFs may be more beneficial than achieving a low prediction error.

Fine-tuning experiments revealed that fine-tuning the F-LSTM layer led to larger performance improvements than fine-tuning the T-LSTM layer. One possible explanation is that frequency information may be more relevant for OVR performance than time information. Alternatively it may indicate that the proposed speech-

dependent augmentation better captures temporal dynamics than spectral characteristics, leaving more room for improvement in the frequency domain. Another explanation could be the larger number of parameters in the F-LSTM layer than in the T-LSTM layer (see Table 3.3), which offers more flexibility during fine-tuning. Training an OVR system only on recorded own voice signals showed limited performance, likely due to the relatively small size of the dataset. This highlights a key advantage of the proposed speech-dependent augmentation technique, which enables training an OVR system when few recorded own voice signals are available. For OVR systems with fewer trainable parameters, a smaller amount of recorded signals may even be sufficient, as investigated in [45].

3.6.1 *Limitations and directions for further research*

While this paper addresses several key aspects of speech-dependent own voice data augmentation for training an OVR system, several limitations remain that may be addressed in future work. First, the experiments were conducted using a relatively limited dataset of recorded own voice signals. Experiments with a larger dataset (including more talkers and utterances, such as the Vibravox dataset [32]) could help determine the amount of recorded own voice signals required to outperform augmentation-based training. Second, our evaluation relies exclusively on objective metrics (PESQ and STOI). Given that the perceived quality of band-limited in-ear own voice signals is hard to predict with standard metrics [52], future work should therefore incorporate subjective listening tests to assess quality improvements more comprehensively. Finally, the current OVR systems are trained as generic systems, aiming to reconstruct own voice of many talkers. It is likely that further performance improvements can be achieved through personalization, e.g., similar to the approach in [53] for personalized speech enhancement.

3.7 Conclusion

In this paper, we have investigated speech-dependent data augmentation techniques for simulating in-ear own voice signals to train an OVR system. Based on a model of own voice transfer characteristics, a large amount of in-ear own voice signals can be simulated from only a limited amount of recorded own voice signals. Experimental results show that speech-dependent individual augmentation yields higher performance than speech-independent and talker-averaged augmentation and outperforms a system directly trained with the recorded own voice signals. Moreover, additional fine-tuning with recorded own voice signals after training with augmented own voice signals, significantly improves performance. When investigating the required amount of recorded own voice signals, it was found that the number of recorded utterances per talker has a smaller influence than the number of recorded talkers. The results show that the proposed speech-dependent individual data augmentation technique for training an OVR system outperforms the EBEN baseline algorithm and is still highly effective even when a limited amount of recorded own voice signals is available.

Abbreviations

OVR: Own voice reconstruction

DNN: Deep neural network

RTF: Relative transfer function

FT-JNF: Frequency and time joint nonlinear filter

STFT: Short-time Fourier transform

WOLA: Weighted overlap-add

LSTM: Long short-term memory

MACs: Multiply-accumulate operations

SNR: Signal-to-noise ratio

PESQ: Perceptual evaluation of speech quality

STOI: Short-time objective intelligibility

EBEN: Extreme bandwidth extension network

Declarations

Availability of data and materials

The datasets used in the current study are available in the Zenodo repositories <https://doi.org/10.5281/zenodo.10844599> (recorded own voice signals) and <https://doi.org/10.5281/zenodo.11196867> (individual transfer functions for simulating spatialized environmental noise). The Common Voice speech dataset used in the current study is available at <https://commonvoice.mozilla.org/en/datasets>, the fifth DNS challenge noise dataset is available through the Github repository <https://github.com/microsoft/DNS-Challenge>. The Python code for the DNN architectures are available in the Github repositories <https://github.com/sp-uhh/deep-non-linear-filter> (FT-JNF) and <https://github.com/jhauret/eben> (EBEN).

Audio examples are available at https://m-ohlenbusch.github.io/own_voice_augmentation_examples/.

Competing interests

Not applicable.

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 Project ID 390895286 – EXC 2177/1.

Authors' contributions

MO developed the algorithms, performed the simulations, analyzed the results, and drafted the article. CR and SD critically discussed the developed algorithms and the simulation results with MO and proofread and revised the article. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

- [1] R. E. Bouserhal, T. H. Falk, and J. Voix, “Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones,” in *Proc. Meetings on Acoustics (ICA)*, vol. 19, Montreal, QC, Canada, 2013. DOI: [10.1121/1.4800452](https://doi.org/10.1121/1.4800452).
- [2] S. Nordholm, A. Davis, P. C. Yong, and H. H. Dam, “Assistive listening headsets for high noise environments: Protection and communication,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Apr. 2015, pp. 5753–5757. DOI: [10.1109/ICASSP.2015.7179074](https://doi.org/10.1109/ICASSP.2015.7179074).
- [3] M. Wang, J. Chen, X.-L. Zhang, and S. Rahardja, “End-to-end multi-modal speech recognition on an air and bone conducted speech corpus,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 513–524, 2023. DOI: [10.1109/TASLP.2022.3224305](https://doi.org/10.1109/TASLP.2022.3224305).
- [4] M. Norda, C. Engel, J. Rennies, J.-E. Appell, S. C. Lange, and A. Hahn, “Evaluating the efficiency of voice control as human machine interface in production,” *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 3, pp. 4817–4828, 2024. DOI: [10.1109/TASE.2023.3302951](https://doi.org/10.1109/TASE.2023.3302951).
- [5] I. López-Espejo, E. Roselló, A. Edraki, N. Harte, and J. Jensen, “Noise-robust hearing aid voice control,” *IEEE Signal Process. Lett.*, vol. 32, pp. 241–245, 2025. DOI: [10.1109/LSP.2024.3512377](https://doi.org/10.1109/LSP.2024.3512377).

- [6] J. Heitkaemper, J. Caroselli, M. McKinnon, A. Narayanan, and N. Howard, “Bone conducted signal guided speech enhancement for voice assistant on earbuds,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025. DOI: [10.1109/ICASSP49660.2025.10889416](https://doi.org/10.1109/ICASSP49660.2025.10889416).
- [7] R. E. Bouserhal, A. Bernier, and J. Voix, “An in-ear speech database in varying conditions of the audio-phonation loop,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 1069–1077, Feb. 2019. DOI: [10.1121/1.5091777](https://doi.org/10.1121/1.5091777).
- [8] B. Gårdbæk and P. Kidmose, “On the origin of cardiovascular sounds recorded from the ear,” *IEEE Trans. Biomed. Eng.*, vol. 72, no. 1, pp. 210–216, 2025. DOI: [10.1109/TBME.2024.3445412](https://doi.org/10.1109/TBME.2024.3445412).
- [9] H. Saint-Gaudens, H. Nélisse, F. Sgard, and O. Doutres, “Towards a practical methodology for assessment of the objective occlusion effect induced by earplugs,” *J. Acoust. Soc. Am.*, vol. 151, no. 6, pp. 4086–4100, Jun. 2022. DOI: [10.1121/10.0011696](https://doi.org/10.1121/10.0011696).
- [10] J. Richard, V. Zimpfer, C. Blondé-Weinmann, and S. Roth, “Change in transfer function between air and bone conduction microphones due to mouth opening variation,” *Applied Acoustics*, vol. 228, p. 110 293, Jan. 2025. DOI: <https://doi.org/10.1016/j.apacoust.2024.110293>.
- [11] S. Stenfelt and S. Reinfeldt, “A model of the occlusion effect with bone-conducted stimulation,” *International Journal of Audiology*, vol. 46, no. 10, pp. 595–608, Jan. 2007. DOI: [10.1080/14992020701545880](https://doi.org/10.1080/14992020701545880).
- [12] S. Vogl and M. Blau, “Individualized prediction of the sound pressure at the eardrum for an earpiece with integrated receivers and microphones,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 917–930, Feb. 2019. DOI: [10.1121/1.5089219](https://doi.org/10.1121/1.5089219).
- [13] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Training strategies for own voice reconstruction in hearing protection devices using an in-ear microphone,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022. DOI: [10.1109/IWAENC53105.2022.9914801](https://doi.org/10.1109/IWAENC53105.2022.9914801).
- [14] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, “Configurable EBEN: Extreme bandwidth extension network to enhance body-conducted speech capture,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3499–3512, 2023. DOI: [10.1109/TASLP.2023.3313433](https://doi.org/10.1109/TASLP.2023.3313433).
- [15] L. He, H. Hou, S. Shi, X. Shuai, and Z. Yan, “Towards bone-conducted vibration speech enhancement on head-mounted wearables,” in *Proc. Annual International Conference on Mobile Systems, Applications and Services*, New York, USA, Jun. 2023, pp. 14–27. DOI: [10.1145/3581791.3596832](https://doi.org/10.1145/3581791.3596832).
- [16] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Modeling of speech-dependent own voice transfer characteristics for hearables with an in-ear microphone,” *Acta Acustica*, vol. 8, 2024. DOI: [10.1051/aacus/2024032](https://doi.org/10.1051/aacus/2024032).

- [17] K. Kondo, T. Fujita, and K. Nakagawa, "On equalization of bone conducted speech for improved speech quality," in *Proc. International Symposium on Signal Processing and Information Technology*, Vancouver, BC, Canada, Aug. 2006, pp. 426–431. DOI: [10.1109/ISSPIT.2006.270839](#).
- [18] M. S. Rahman and T. Shimamura, "Intelligibility enhancement of bone conducted speech by an analysis-synthesis method," in *Proc. International Midwest Symposium on Circuits and Systems (MWSCAS)*, Seoul, South Korea, Aug. 2011. DOI: [10.1109/MWSCAS.2011.6026374](#).
- [19] H. S. Shin, T. Fingscheidt, and H.-G. Kang, "A priori SNR estimation using air- and bone-conduction microphones," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 11, pp. 2015–2025, Nov. 2015. DOI: [10.1109/TASLP.2015.2446202](#).
- [20] R. E. Bouserhal, T. H. Falk, and J. Voix, "In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension," *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 1321–1331, Mar. 2017. DOI: [10.1121/1.4976051](#).
- [21] H.-P. Liu, Y. Tsao, and C.-S. Fuh, "Bone-conducted speech enhancement using deep denoising autoencoder," *Speech Communication*, vol. 104, pp. 106–112, Nov. 2018. DOI: [10.1016/j.specom.2018.06.002](#).
- [22] H. Park, Y.-S. Shin, and S.-H. Shin, "Speech quality enhancement for in-ear microphone based on neural network," *IEICE Trans. on Information and Systems*, vol. 102, no. 8, pp. 1594–1597, 2019. DOI: [10.1587/transinf.2018EDL8249](#).
- [23] C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1888–1900, 2020. DOI: [10.1109/TASLP.2020.2976193](#).
- [24] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, "Time-domain multi-modal bone/air conducted speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 1035–1039, 2020. DOI: [10.1109/LSP.2020.3000968](#).
- [25] M. Wang, J. Chen, X. Zhang, Z. Huang, and S. Rahardja, "Multi-modal speech enhancement with bone-conducted speech in time domain," *Applied Acoustics*, vol. 200, no. 109058, Nov. 2022. DOI: [10.1016/j.apacoust.2022.109058](#).
- [26] H. Wang, X. Zhang, and D. Wang, "Attention-based fusion for bone-conducted and air-conducted speech enhancement in the complex domain," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7757–7761. DOI: [10.1109/ICASSP43922.2022.9746374](#).
- [27] Y. Li, Y. Wang, X. Liu, Y. Shi, S. Patel, and S.-F. Shih, "Enabling real-time on-chip audio super resolution for bone-conduction microphones," *Sensors*, vol. 23, no. 1, Jan. 2023. DOI: [10.3390/s23010035](#).

- [28] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Multi-microphone noise data augmentation for DNN-based own voice reconstruction for hearables in noisy environments,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 416–420. DOI: 10.1109/ICASSP48485.2024.10447066.
- [29] C. Li, F. Yang, and J. Yang, “Restoration of bone-conducted speech with U-Net-like model and energy distance loss,” *IEEE Signal Process. Lett.*, vol. 31, pp. 166–170, 2024. DOI: 10.1109/LSP.2023.3347149.
- [30] C. Li, F. Yang, and J. Yang, “A two-stage approach to quality restoration of bone-conducted speech,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 818–829, 2024. DOI: 10.1109/TASLP.2023.3337988.
- [31] X. Hu, Z. Chen, and F. Yin, “Bone-conducted speech codec based on AMR-WB framework and MHSA-CycleGAN network,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 33, pp. 124–138, 2025. DOI: 10.1109/TASLP.2024.3511252.
- [32] J. Hauret, M. Olivier, T. Joubaud, C. Langrenne, S. Poirée, V. Zimpfer, and É. Bavu, “Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors,” *Speech Communication*, Apr. 2025. DOI: 10.1016/j.specom.2025.103238.
- [33] H. Wang, X. Zhang, and D. Wang, “Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 3134–3143, 2022. DOI: 10.1109/TASLP.2022.3209943.
- [34] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, “Speaker Adaptation For Enhancement Of Bone-Conducted Speech,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 10 456–10 460. DOI: 10.1109/ICASSP48485.2024.10447322. Accessed: Mar. 26, 2024.
- [35] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 351–355. DOI: 10.1109/ICASSP.2018.8461310.
- [36] P. Masztalski, M. Matuszewski, K. Piaskowski, and M. Romaniuk, “StoRIR: Stochastic Room Impulse Response Generation for Audio Data Augmentation,” in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 2857–2861. DOI: 10.21437/Interspeech.2020-2261.
- [37] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, “gpuRIR: A python library for room impulse response simulation with GPU acceleration,” *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021. DOI: 10.1007/s11042-020-09905-3.

- [38] F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J.-H. Bach, J. Wellmann, and B. Kollmeier, “A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research,” in *Proc. AES International Conference on Headphone Technology*, San Francisco, USA, Aug. 2019.
- [39] K. Tesch and T. Gerkmann, “Insights into deep non-linear filters for improved multi-channel speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 563–575, 2023. DOI: 10.1109/TASLP.2022.3221046.
- [40] Y. Avargel and I. Cohen, “On multiplicative transfer function approximation in the short-time Fourier transform domain,” *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007. DOI: 10.1109/LSP.2006.888292.
- [41] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR Corpus Based on Public Domain Audio Books,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.
- [42] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common Voice: A massively-multilingual speech corpus,” in *Proc. 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4218–4222.
- [43] R. Huber, A. Pusch, N. Moritz, J. RENNIES, H. Schepker, and B. T. Meyer, “Objective assessment of a speech enhancement scheme with an automatic speech recognition-based system,” in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 86–90.
- [44] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh, and R. Aichner, “ICASSP 2023 Deep Noise Suppression Challenge,” *IEEE Open Journal of Signal Processing*, vol. 5, pp. 725–737, Mar. 2024. DOI: 10.1109/OJSP.2024.3378602.
- [45] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Low-complexity own voice reconstruction for hearables with an in-ear microphone,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025. DOI: 10.1109/ICASSP49660.2025.10887874.
- [46] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, “STFT-domain neural speech enhancement with very low algorithmic latency,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 397–410, 2023. DOI: 10.1109/TASLP.2022.3224285.
- [47] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, “Towards efficient models for real-time deep noise suppression,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 656–660. DOI: 10.1109/ICASSP39728.2021.9413580.
- [48] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. International Conference on Learning Representations*, San Diego, USA, 2015. DOI: 10.48550/arXiv.1412.6980.

- [49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [50] 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,” *International Telecommunications Union*, Feb. 2001.
- [51] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011. DOI: [10.1109/TASL.2011.2114881](https://doi.org/10.1109/TASL.2011.2114881).
- [52] J. Richard, V. Zimpfer, and S. Roth, “Comparison of objective and subjective methods for evaluating speech quality and intelligibility recorded through bone conduction and in-ear microphones,” *Applied Acoustics*, vol. 211, Aug. 2023. DOI: [10.1016/j.apacoust.2023.109576](https://doi.org/10.1016/j.apacoust.2023.109576). Accessed: Oct. 10, 2023.
- [53] A. Kuznetsova, A. Sivaraman, and M. Kim, “The potential of neural speech synthesis-based data augmentation for personalized speech enhancement,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 2023. DOI: [10.1109/ICASSP49357.2023.10096601](https://doi.org/10.1109/ICASSP49357.2023.10096601).

LOW-COMPLEXITY OWN VOICE RECONSTRUCTION FOR HEARABLES WITH AN IN-EAR MICROPHONE

This chapter is identical in content to the publication: M. Ohlenbusch, C. Rollwage, and S. Doclo, “Low-complexity own voice reconstruction for hearables with an in-ear microphone,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025. DOI: 10.1109/ICASSP49660.2025.10887874.

Authors	Author's contribution							
	A	B	C	D	E	F	G	H
Mattes Ohlenbusch	✗	✗	✗	✗	✗	✗	✗	✗
Christian Rollwage	✗						✗	✗
Simon Doclo	✗					✗	✗	✗

A - Substantial contributions to the conception or design of the work

B - Acquisition of the data

C - Analysis of the data

D - Interpretation of the data

E - Drafting the work

F - Revising the work critically

G - Final approval of the version to be published

H - Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Abstract

Hearable devices, equipped with one or more microphones, are commonly used for speech communication. Here, we consider the scenario where a hearable is used to capture the user's own voice in a noisy environment. In this scenario, own voice reconstruction (OVR) is essential for enhancing the quality and intelligibility of the recorded noisy own voice signals for telephony applications. In previous work, we developed a deep learning-based OVR system, aiming to reduce the amount of device-specific recorded signals for training by using data augmentation with phoneme-dependent models of own voice transfer characteristics. Given the limited computational resources available on hearables, in this paper we propose low-complexity variants of an OVR system based on the frequency-and-time joint non-linear filter (FT-JNF) architecture and investigate the required amount of device-specific recorded signals for effective data augmentation and fine-tuning. Simulation results show that the proposed OVR system considerably improves speech quality, even under constraints of low complexity and a limited amount of device-specific recorded signals.

4.1 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. 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 [1]. The goal of own voice reconstruction (OVR) is to estimate clean broadband own voice signals from the outer and/or in-ear microphone signals.

Several OVR approaches have been proposed which extend the bandwidth of the in-ear microphone signal [2–7]¹. However, it has been shown in [8–12] that speech quality can be further improved by using outer microphones in addition to in-ear microphones. Although previously proposed deep learning-based OVR systems often have high computational complexity and millions of parameters, it is crucial that OVR systems for hearables have low complexity and few parameters to meet hardware requirements. In addition, training an OVR system typically requires a large amount of device-specific own voice signals. Whereas some OVR approaches only use device-specific recorded own voice signals directly as training data, e.g., [10, 13], other approaches perform training with augmented own voice data generated from a small amount of device-specific recorded signals and then perform fine-tuning with the recorded own voice signals [2, 4, 5, 12]. For single-channel speech enhancement systems, the amount of required training data tends to decrease as complexity decreases [14]. However, it is unclear if this relationship also applies to training low-complexity OVR systems with augmented own voice data and fine-tuning using only few device-specific own voice recorded signals.

In this paper, we propose low-complexity variants of an OVR system based on the frequency-and-time joint non-linear filter (FT-JNF) architecture [15]. We train the OVR system variants using a phoneme-dependent own voice data augmentation method proposed in [12]. We compare the OVR performance of the proposed system variants, differing in size and computational complexity, with baseline systems. In addition, we investigate the influence of the amount of device-specific recorded own voice signals used for data augmentation and fine-tuning on the OVR performance. Experimental results show that the proposed system outperforms baseline systems at a comparable complexity, even when only a small amount of device specific recorded signals is available.

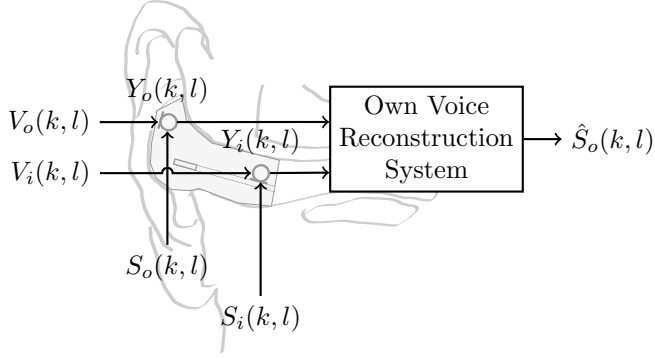


Figure 4.1: Block diagram of own voice reconstruction using an outer and an in-ear microphone of a hearable.

4.2 Signal model

We consider a hearable device equipped with an outer microphone and an in-ear microphone, as depicted in Fig. 4.1. The signals are denoted by subscripts o for the outer microphone and i for the in-ear microphone. In the short-time Fourier transform (STFT) domain, $S_o(k, l)$ and $S_i(k, l)$ denote the own voice signals of the user at both microphones, where k and l denote the frequency index and the frame index. The outer and in-ear microphone signals are given by

$$Y_o(k, l) = S_o(k, l) + V_o(k, l), \quad (4.1)$$

$$Y_i(k, l) = S_i(k, l) + V_i(k, l), \quad (4.2)$$

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

4.3 Own voice reconstruction system

The goal of own voice reconstruction is to estimate the own voice signal $S_o(k, l)$ from the outer and in-ear microphone signals. It is assumed here that this signal is spectrally similar to the own voice at the talker's mouth. In [11, 12] an OVR system based on the FT-JNF architecture [15] has been proposed, see Fig. 4.2. This system takes the complex-valued outer and in-ear microphone STFT coefficients as input, split into real and imaginary parts $Y_o^{\text{re}}(k, l)$ and $Y_o^{\text{im}}(k, l)$ for the outer microphone and $Y_i^{\text{re}}(k, l)$ and $Y_i^{\text{im}}(k, l)$ for the in-ear microphone. The input is processed by a frequency-direction LSTM (F-LSTM) with H_f hidden units, followed by a time-

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

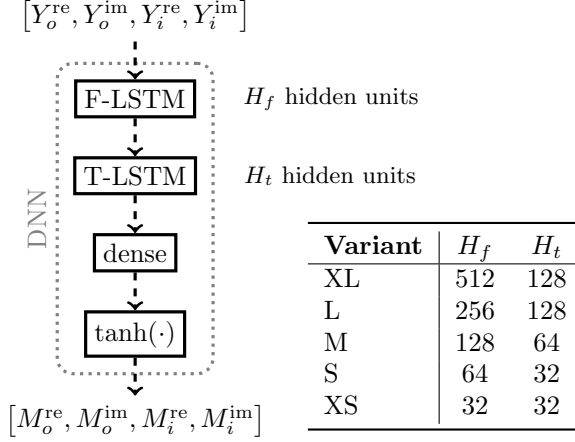


Figure 4.2: Architecture of the OVR system, estimating complex-valued masks for the outer and in-ear microphones. The number of hidden units, H_f and H_t , is different for each proposed variant.

direction LSTM (T-LSTM) with H_t hidden units. The output of the T-LSTM layer is then processed by a dense layer and a tanh activation function to obtain the real and imaginary parts of the complex-valued STFT masks $M_o(k, l)$ and $M_i(k, l)$ for the outer and in-ear microphones. The tanh activation function constrains the real and imaginary parts of the masks to the range $[-1, 1]$. To compute the own voice estimate $\hat{S}_o(k, l)$, noisy STFT coefficients of both microphones are multiplied with the corresponding masks and added, i.e.,

$$\hat{S}_o(k, l) = \sum_{m \in \{o, i\}} M_m(k, l) \cdot Y_m(k, l). \quad (4.3)$$

In this paper we consider several variants of this OVR system, which differ in size and computational complexity. We vary the size by changing the number of hidden units H_f and H_t , see numbers in Fig. 4.2 for the extra-large (XL), large (L), medium (M), small (S) and extra-small (XS) variants. In this paper, we do not consider further complexity reduction, e.g., by quantization or pruning [16].

4.4 Phoneme-dependent own voice augmentation

Training an OVR system using both an outer and an in-ear microphone requires a device-specific dataset of own voice signals. Recording a dataset sufficiently large for direct training requires considerable effort. However, it is more feasible to record only a small dataset with a limited number of talkers or utterances per talker. In [17], a method was proposed to simulate in-ear own voice signals from outer microphone signals (for talkers and utterances not included in the small recorded dataset). From the small recorded dataset, phoneme-specific relative transfer functions (RTFs) be-

tween the outer and the in-ear microphone are estimated for each recorded talker, over all time frames in which a specific phoneme p occurs. The estimated phoneme-specific RTFs are denoted by $\hat{H}_p(k)$. For RTF estimation, it is assumed here that there is no environmental noise in the small recorded dataset, and that the sensor noise is negligible compared to the own voice in both microphone signals.

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

$$\hat{S}_i(k, l) = \hat{H}_{p_o(l)}(k) \cdot S_o(k, l). \quad (4.4)$$

Additionally, to avoid artifacts during phoneme transitions, temporal smoothing of $\hat{H}_{p_o(l)}(k)$ is carried out (see [17] for details). Instead of assuming recorded outer microphone signals are available, it was proposed in [12] to use clean speech signals from standard datasets instead. Since standard datasets are readily available, this method allows for the simulation of a large amount of simulated in-ear own voice signals. 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. In [12] it was shown that when a small dataset of device-specific recorded signals is available, this data augmentation method can improve OVR performance compared to only using the device-specific recorded signals directly as training data. 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.

4.5 Experimental setup

To evaluate the proposed FT-JNF variants and several baseline systems (see Section 4.5.4) for own voice reconstruction, we conduct an experimental evaluation. In this section, we describe the experimental setup for the evaluation.

4.5.1 Datasets

The evaluation uses clean recorded own voice signals made with the Hearpiece hearable prototype [18]. A dataset of German own voice signals of 18 talkers with 306 utterances each is split into training, validation, and test sets with 12, 2, and 4 talkers, respectively. All OVR systems are first trained on augmented own voice signals and then fine-tuned on recorded own voice signals. The augmented own voice signals are obtained by augmenting 10% of the German portion of the CommonVoice dataset [19] (v11.0), corresponding to 115.7 hours, as described in Section 4.4. The full augmented training and fine-tuning of the proposed variants and the baseline systems uses recorded signals from 12 talkers with 306 utterances each. Reduced amounts are considered in Section 4.6.2. It should be noted that independent of the amount of used device-specific recorded signals, all systems were trained with the same amount of augmented data (115.7 hours).

The noise signals at both microphones used for training and testing are a spatialized version of the fifth DNS challenge [20], obtained following the procedure in [11] using individually matched, measured transfer functions for the same users as in the dataset of recorded own voice signals². Measurements from 8 horizontal directions in 45°-steps with 1.5 m distance are used to compute either point source signals (single direction) or pseudo-diffuse noise signals (8 directions).

4.5.2 Training details

The experiments are conducted at a sampling rate of 16 kHz, using an STFT framework with a frame length of 32 ms and a frame shift of 16 ms, where a square-root Hann window is used both in analysis and synthesis. Own voice and noise signals are mixed at a random signal-to-noise ratio (SNR) between -10 and 25 dB, defined at the outer microphone. Training is carried out with four examples per batch and an example length of 3 s, using the combined L_1 loss between the target clean own voice signal at the outer microphone and the estimated own voice signal in the time domain and the STFT domain (after re-analysis) [21]. The ADAM optimizer [22] is used with an initial learning rate of 10^{-4} , which is halved after three epochs without improvement of the validation loss, and training is stopped after six epochs without improvement. The initial learning rate for fine-tuning is 10^{-5} .

4.5.3 Evaluation metrics

OVR performance is evaluated using wideband PESQ [23], extended short-time objective intelligibility (ESTOI) [24], and log-spectral distance (LSD) [25]. For all three metrics, the clean own voice signal at the outer microphone is chosen as the reference signal. Higher PESQ and ESTOI values are better, while lower LSD values are better. During testing, OVR performance is evaluated at SNRs of -10, -5, 0, 5, and 10 dB. The results are averaged over the test set and over SNR. System complexity is reported in terms of number of parameters, number of multiply-accumulate operations per second (MACs/s), and real-time factor (RF). MACs are computed using the `thop` Python package. The RF is computed on an Intel Core i7-10850H CPU (2.7 GHz).

4.5.4 Baseline systems

The baseline systems include three systems that only use the in-ear microphone (IM) signal, and one system that uses both the outer and in-ear microphone signals. All baseline systems were retrained using the same setup as described in Section 4.5.2 for the proposed FT-JNF variants:

- UNet [2, 26]: Time-domain system performing reconstruction of the in-ear own voice signal.

² German own voice dataset [online]: <https://doi.org/10.5281/zenodo.10844599>, individual transfer function measurements [online]: <https://doi.org/10.5281/zenodo.11196867>

Table 4.1: Performance, size and complexity of the baseline systems and the proposed FT-JNF variants (XL, L, M, S, XS). 'M' indicates millions, 'G' indicates billions. Rows with a gray background indicate systems using only the in-ear microphone.

	Intrusive metrics			Size and complexity		
System	PESQ	ESTOI	LSD	Param.	MACs/s	RF
Unprocessed	1.25	0.51	2.46	-	-	-
UNet (IM) [2]	1.85	0.65	1.30	10.278 M	6.03 G	0.157
EBEN (IM) [3]	1.51	0.57	1.64	1.946 M	1.02 G	0.034
FT-JNF XL (IM)	1.47	0.61	1.73	1.390 M	22.38 G	0.387
GCBFSNet [27]	1.93	0.68	1.36	0.100 M	0.31 G	0.303
FT-JNF XL	2.58	0.78	1.08	1.390 M	22.45 G	0.392
FT-JNF L	2.50	0.77	1.10	0.466 M	7.55 G	0.173
FT-JNF M	2.22	0.72	1.27	0.118 M	1.93 G	0.071
FT-JNF S	2.18	0.72	1.28	0.031 M	0.50 G	0.029
FT-JNF XS	1.95	0.68	1.40	0.013 M	0.23 G	0.011

- Extreme Bandwidth Extension Network (EBEN) [3]: Time-domain system, originally proposed for bandwidth extension of body-conducted speech. The generator was retrained by replacing the generative adversarial network-based training with the loss function from [21], as used for all other systems.
- FT-JNF XL (IM): The proposed FT-JNF XL using only the in-ear microphone signal. Due to the activation function and only estimating masks (see Section 4.3), this system is unable to perform bandwidth extension.
- group communication binaural filter-and-sum network (GCBFSNet) [27]: Unilateral version of the low-complexity GCBFSNet (8 groups, 32 hidden units, with post-filter, 2 ms frames, 1 ms frame shift), retrained for OVR using both the outer and in-ear microphone signals.

4.6 Results

In this section, the results of the experimental evaluation are presented. In Section 4.6.1, the proposed FT-JNF variants are compared to the baseline systems. In Section 4.6.2, the influence of the amount of device-specific recorded own voice signals for data augmentation and fine-tuning is investigated. Audio examples from the evaluation are available online³.

4.6.1 Comparison to baseline systems

Table 4.1 compares the performance, size and complexity of the proposed FT-JNF variants and the baseline systems. First, it can be observed that all OVR variants achieve considerable improvements in all metrics compared to the unprocessed (noisy outer microphone) signals. Not surprisingly, systems using both the outer microphone and the in-ear microphone (GCBFSNet and FT-JNF variants) generally outperform systems using only the in-ear microphone (UNet, EBEN, FT-JNF XL (IM)). Among the systems using only the in-ear microphone, UNet achieves the best scores but also has the most parameters. While EBEN and FT-JNF XL (IM) have a similar amount of parameters and performance, FT-JNF XL (IM) has a much higher complexity (MACs/s and RF). Among the systems using both the outer and the in-ear microphone, GCBFSNet has a slightly lower RF than the FT-JNF XL variant, but higher than the L, M, S, XS variants. Although GCBFSNet has fewer MACs/s than FT-JNF S, FT-JNF S has about three times fewer parameters and achieves better scores in all metrics. FT-JNF XS performs comparable to GCBFSNet with fewer MACs/s and at a much lower RF.

FT-JNF XL performs much better than IM-FT-JNF XL, while the complexity of FT-JNF XL is only marginally higher. This indicates a substantial performance gain from using the outer microphone. Due to performing masking in a constrained value range, FT-JNF XL (IM) is unable to reconstruct speech in high frequency regions, whereas FT-JNF XL can use high frequency content from the outer microphone.

While FT-JNF XL consists of 1.39 million parameters, it requires a high number of computations due to the F-LSTM iterating over all frequencies for each time frame. When the model complexity is decreased to L, the MACs/s and RF also decrease, while the performance only slightly decreases. Even though the performance of the smaller variants (M, S and XS) is lower compared to FT-JNF XL and L, their performance is still better than the baseline systems. It should be noted that FT-JNF S and XS require approximately 44 and 97 times fewer MACs/s than FT-JNF XL, respectively.

4.6.2 Influence of amount of device-specific recorded signals

To investigate the relationship between system complexity and amount of device-specific recorded signals, the baseline systems and the proposed FT-JNF variants were retrained using different amounts of device-specific recorded signals for augmented training and fine-tuning. We investigated both the influence of reducing the number of talkers from 12 to 3 (with 306 utterances) and reducing the number of utterances from 306 to 25 (for 12 talkers). Fig. 4.3 shows the results in terms of PESQ improvement (Δ PESQ) compared to the unprocessed noisy outer microphone signals. When the number of talkers is reduced, the performance of baselines with low complexity (GCBFSNet, EBEN) only slightly decreases while for UNet and FT-JNF XL (IM) there is a larger decrease. For the proposed variants, a large performance decrease from a reduced number of talkers is observed for the XL and L

³ Audio examples [online]: https://m-ohlenbusch.github.io/low_complexity_ovr_examples/

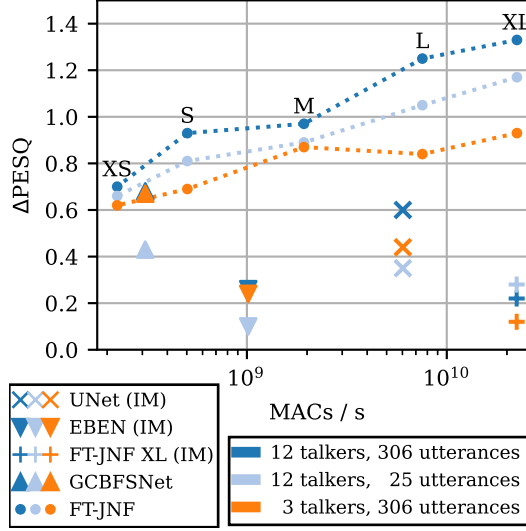


Figure 4.3: PESQ improvement of the baseline systems and the proposed FT-JNF variants for different amounts of device-specific recorded signals (talkers, utterances). Different systems are distinguished by different symbols, while different amounts of recorded signals are represented by different colors.

variants, while the performance only slightly decreases for the M, S, and XS variants. When the number of recorded utterances per talker is reduced, for the baselines the performance decrease is larger than when reducing the number of talkers, but it is smaller for the proposed variants. The results indicate that low-complexity OVR systems require fewer device-specific recorded signals for augmented training and fine-tuning than systems with higher computational complexity.

4.7 Conclusion

In this paper, we proposed variants of the FT-JNF architecture with low computational complexity for OVR. We investigated the influence of the amount of device-specific recorded signals used for data augmentation and fine-tuning on the OVR performance. Experimental results demonstrate that the proposed variants outperform baseline systems at a comparable complexity. Even under constraints of low complexity and a limited amount of device-specific recorded signals available for training, considerable quality improvements can be achieved by the proposed system.

Acknowledgement

The authors would like to thank Nils L. Westhausen for providing the code for the original GCBFSNet architecture.

References

- [1] R. E. Bouserhal, A. Bernier, and J. Voix, “An In-Ear Speech Database in Varying Conditions of the Audio-Phonation Loop,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 1069–1077, Feb. 2019.
- [2] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Training Strategies for Own Voice Reconstruction in Hearing Protection Devices Using An In-Ear Microphone,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022.
- [3] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, “Configurable EBEN: Extreme Bandwidth Extension Network to Enhance Body-conducted Speech Capture,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 3499–3512, 2023.
- [4] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, “Speaker Adaptation For Enhancement Of Bone-Conducted Speech,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 10 456–10 460. Accessed: Mar. 26, 2024.
- [5] Y. Sui, M. Zhao, J. Xia, X. Jiang, and 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,” *arXiv:2405.01242v3*, May 2024.
- [6] C. Li, F. Yang, and J. Yang, “Restoration of Bone-Conducted Speech with U-Net-like Model and Energy Distance Loss,” *IEEE Signal Processing Letters*, vol. 31, pp. 166–170, 2024. Accessed: Jan. 9, 2024.
- [7] C. Li, F. Yang, and J. Yang, “A Two-Stage Approach to Quality Restoration of Bone-Conducted Speech,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 818–829, 2024. Accessed: Jan. 9, 2024.
- [8] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, “Time-Domain Multi-Modal Bone/Air Conducted Speech Enhancement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.
- [9] M. Wang, J. Chen, X. Zhang, Z. Huang, and S. Rahardja, “Multi-Modal Speech Enhancement with Bone-Conducted Speech in Time Domain,” *Applied Acoustics*, vol. 200, no. 109058, Nov. 2022. Accessed: Feb. 28, 2023.
- [10] H. Wang, X. Zhang, and D. Wang, “Fusing Bone-Conduction and Air-Conduction Sensors for Complex-Domain Speech Enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 3134–3143, 2022.
- [11] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Multi-Microphone Noise Data Augmentation for DNN-based Own Voice Reconstruction for Hearables in Noisy Environments,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 416–420. Accessed: Dec. 18, 2023.

- [12] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Speech-dependent Data Augmentation for Own Voice Reconstruction with Hearable Microphones in Noisy Environments,” *arXiv:2405.11592*, May 2024.
- [13] J. Hauret, M. Olivier, T. Joubaud, C. Langrenne, S. Poirée, V. Zimpfer, and É. Bavu, “Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors,” *arXiv:2407.11828*, Jul. 2024.
- [14] W. Zhang, K. Saijo, J.-w. Jung, C. Li, S. Watanabe, and Y. Qian, “Beyond Performance Plateaus: A Comprehensive Study on Scalability in Speech Enhancement,” in *Proc. Interspeech*, Kos, Greece, Sep. 2024, pp. 1740–1744.
- [15] K. Tesch and T. Gerkmann, “Insights Into Deep Non-Linear Filters for Improved Multi-Channel Speech Enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023.
- [16] M. Stamenovic, N. L. Westhausen, L.-C. Yang, C. Jensen, and A. Pawlicki, “Weight, Block or Unit? Exploring Sparsity Tradeoffs for Speech Enhancement on Tiny Neural Accelerators,” in *Proc. NeurIPS Workshop Efficient Natural Language and Speech Processing*, Nov. 2021.
- [17] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Modeling of speech-dependent own voice transfer characteristics for hearables with an in-ear microphone,” *Acta Acustica*, vol. 8, 2024.
- [18] F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J.-H. Bach, J. Wellmann, and B. Kollmeier, “A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research,” in *Proc. AES International Conference on Headphone Technology*, San Francisco, USA, Aug. 2019.
- [19] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common Voice: A Massively-Multilingual Speech Corpus,” in *Proc. 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4218–4222.
- [20] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh, and R. Aichner, “ICASSP 2023 Deep Noise Suppression Challenge,” *IEEE Open Journal of Signal Processing*, vol. 5, pp. 725–737, 2024. Accessed: Mar. 27, 2024.
- [21] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, “STFT-domain neural speech enhancement with very low algorithmic latency,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 397–410, 2023.
- [22] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. International Conference on Learning Representations*, San Diego, USA, 2015.
- [23] 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,” *International Telecommunications Union*, Feb. 2001, Geneva, Switzerland.

- [24] J. Jensen and C. H. Taal, “An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
- [25] A. Gray and J. Markel, “Distance Measures for Speech Processing,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.
- [26] H. Wang and D. Wang, “Towards Robust Speech Super-Resolution,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 2058–2066, 2021.
- [27] 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, Dec. 2023.

SUBJECTIVE QUALITY EVALUATION OF PERSONALIZED OWN VOICE RECONSTRUCTION SYSTEMS

This chapter is identical in content to the publication: M. Ohlenbusch, C. Rollwage, S. Doclo, and J. Rennies, “Subjective quality evaluation of personalized own voice reconstruction systems,” *Acta Acustica*, vol. 10, no. 26, 2026. DOI: [10.1051/aacus/2026021](https://doi.org/10.1051/aacus/2026021).

Authors	Author's contribution							
	A	B	C	D	E	F	G	H
Mattes Ohlenbusch	✗	✗	✗	✗	✗	✗	✗	✗
Christian Rollwage	✗						✗	✗
Simon Doclo	✗					✗	✗	✗
Jan Rennies	✗			✗		✗	✗	✗

A - Substantial contributions to the conception or design of the work

B - Acquisition of the data

C - Analysis of the data

D - Interpretation of the data

E - Drafting the work

F - Revising the work critically

G - Final approval of the version to be published

H - Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

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.

5.1 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]¹, which is achieved by training them with data from multiple talkers in order to achieve robustness to 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

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

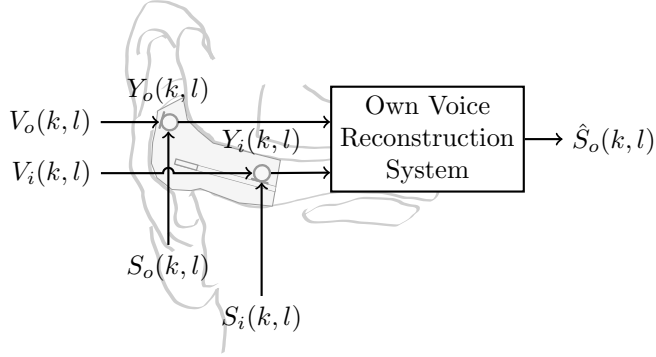


Figure 5.1: Block diagram of own voice reconstruction using an outer and an in-ear microphone of a hearable.

(without pre-training). While in both [21] and [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.

5.2 Own voice reconstruction

5.2.1 Signal model

We consider a hearable device equipped with an outer microphone and an in-ear microphone, as depicted in Fig. 5.1. 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 index. The noisy outer and in-ear microphone signals are given by

$$Y_o(k, l) = S_o(k, l) + V_o(k, l), \quad (5.1)$$

$$Y_i(k, l) = S_i(k, l) + V_i(k, l), \quad (5.2)$$

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.

5.2.2 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.,

$$\hat{S}_o(k, l) = \mathcal{D}\{Y_o(k, l), Y_i(k, l)\}. \quad (5.3)$$

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.,

$$\hat{S}_o(k, l) = \bar{\mathcal{D}}\{Y_o^a(k, l), Y_i^a(k, l)\}. \quad (5.4)$$

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.,

$$\hat{S}_o^a(k, l) = \mathcal{D}^a\{Y_o^a(k, l), Y_i^a(k, l)\}. \quad (5.5)$$

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 5.2 shows the OVR system architecture considered in this work. Different

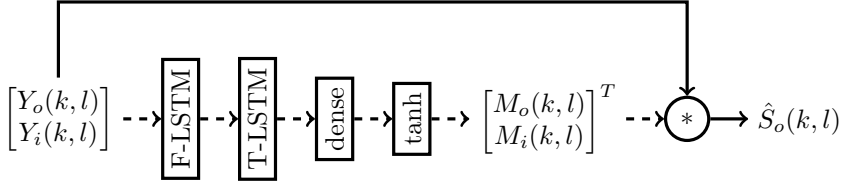


Figure 5.2: DNN-based own voice reconstruction system based on the FT-JNF architecture from [24]. The first LSTM layer operates in frequency direction (F-LSTM) and the second LSTM layer operates in time direction (T-LSTM). Figure adapted from [18].

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

$$\hat{S}_o(k, l) = \sum_{m \in \{o, i\}} M_m(k, l) \cdot Y_m(k, l). \quad (5.6)$$

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

5.2.3 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.

5.2.3.1 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

$$\hat{S}_i(k, l) = \hat{H}_{p_o(l)}(k) \times S_o(k, l), \quad (5.7)$$

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.

5.2.3.2 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 (5.7), the simulated in-ear signal $\hat{S}_i^a(k, l)$ of target talker a is obtained as

$$\hat{S}_i^a(k, l) = \hat{H}_{p_o(l)}^a(k) \times S_o(k, l). \quad (5.8)$$

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.

5.2.3.3 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 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.

5.3 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. 5.3.1), and in addition to generic data augmentation and fine-tuning, personalized data augmentation and fine-tuning are also considered.

5.3.1 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 Common-Voice 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.

5.3.2 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. 5.4). 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.

5.4 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 [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^q The generalized power spectrum model for quality (GPSM^q) [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 efficient model for binaural audio quality (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

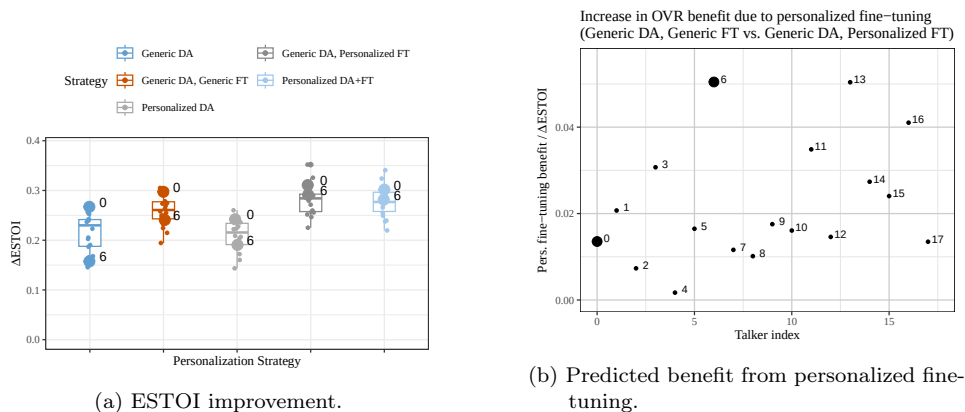


Figure 5.3: (a) ESTOI improvement achieved by OVR systems trained with different personalization strategies in data augmentation (DA) and fine-tuning (FT) and (b) individual difference in ESTOI improvement between the conditions Generic DA, generic FT and generic DA, personalized FT. Individual data 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.

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].

5.5 Results of instrumental assessment

Figure 5.3a and Figure 5.4a show the improvements (Δ) 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 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

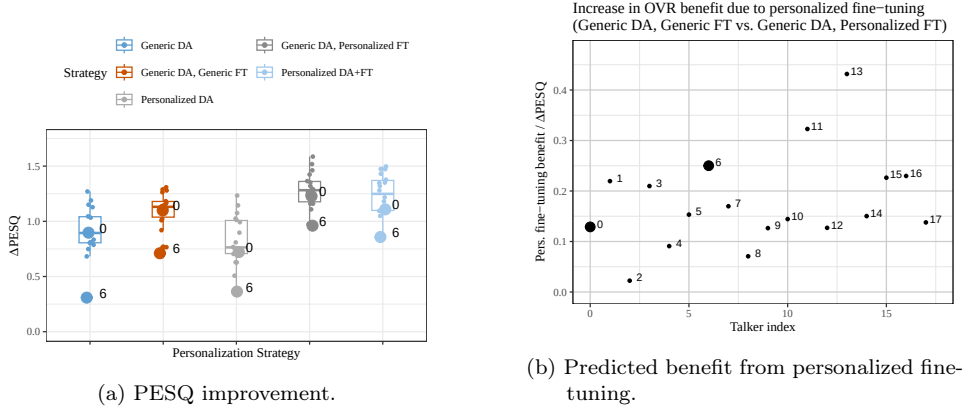


Figure 5.4: (a) PESQ improvement achieved by OVR systems trained with different personalization strategies in data augmentation (DA) and fine-tuning (FT) and (b) individual difference in PESQ improvement between the conditions generic DA, generic FT and generic DA, personalized FT. Individual data points 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.

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 Figs. 5.3 and 5.4, two example target talkers are highlighted. For assessing the benefit of personalized over generic fine-tuning, the difference between the Δ ESTOI and Δ PESQ scores in Figs. 5.3a and 5.4a for the conditions Generic DA, Generic FT and Generic DA, personalized FT are shown in Figs. 5.3b and 5.4b. 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 Δ 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.

5.6 Listening experiment

5.6.1 *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. 5.6.3) than the lowpass-filtered clean speech signals defined in [50]. The experiment was carried out using the Web-MUSHRA 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).

5.6.2 *Participants*

Twenty-five normal-hearing native German-speaking participants (13 female, 12 male), aged 23.8 ± 3.8 years (mean $\pm$ standard deviation), took part in the listening experiment. All participants had pure-tone thresholds ≤ 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).

5.6.3 *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 Sect. 5.5 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 Sect. 5.5. 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. 5.3.1, 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 Sect. 5.3.1. 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 Fig. 5.5. 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

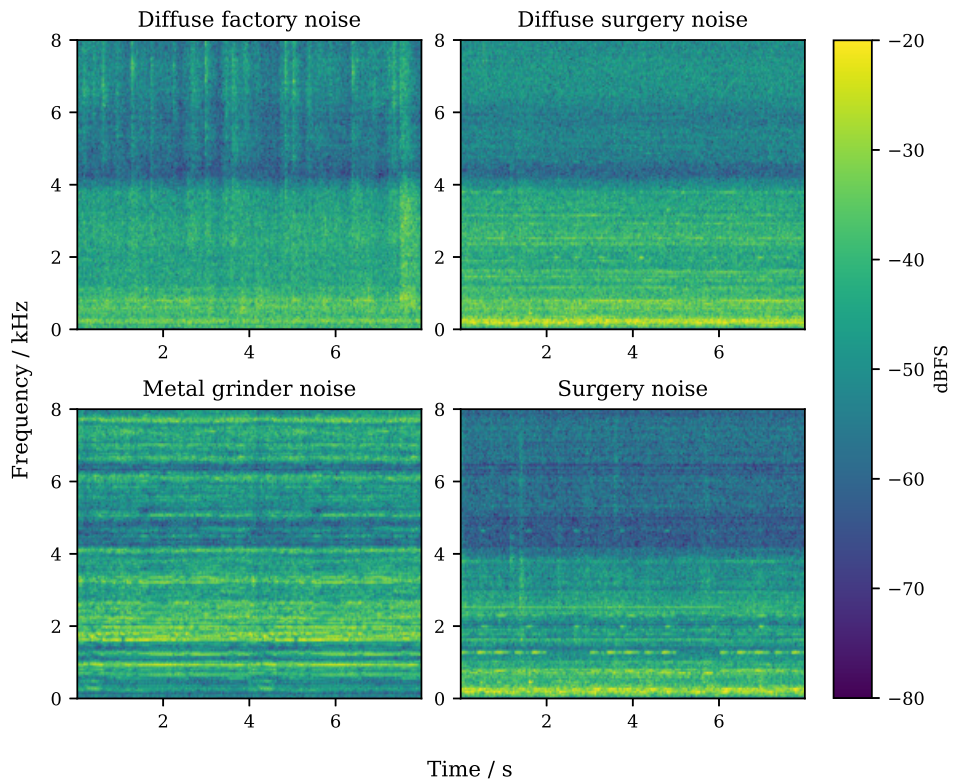


Figure 5.5: Spectrograms of the recorded noise signals used in the subjective evaluation.

condition, talker, and noise type, utterances of three different sentences were presented to the participants, and the resulting subjective ratings were averaged over utterances.

5.7 Results of listening experiment

5.7.1 Subjective quality ratings

Figure 5.6 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 5.7 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 Sect. 5.5 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.

5.7.2 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-

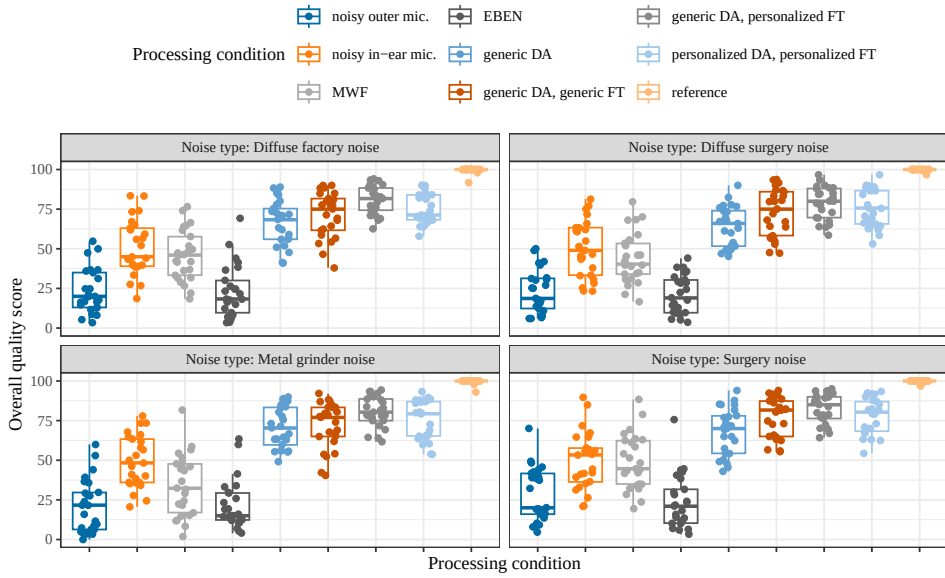


Figure 5.6: Subjective MUSHRA quality ratings (averaged over sentences) for speech in the *low predicted benefit* case.

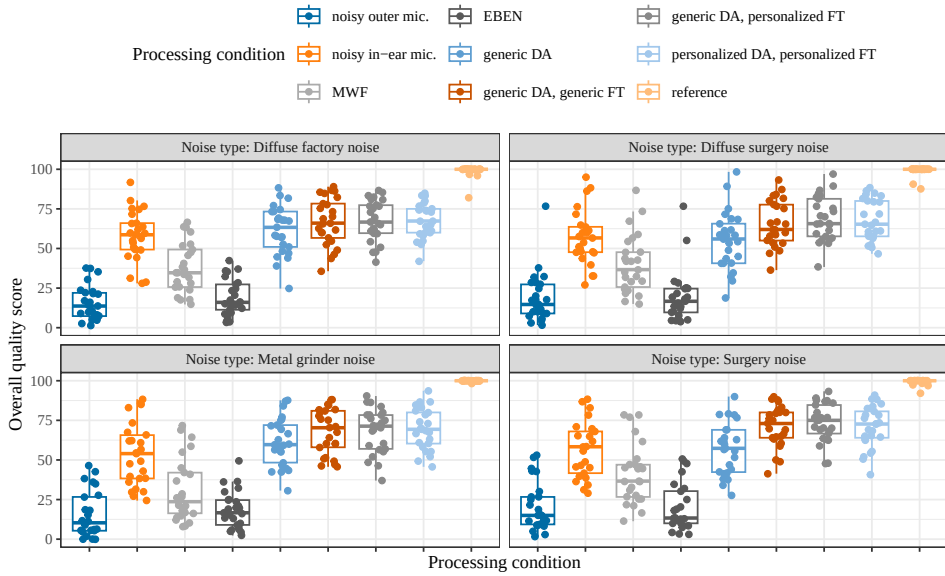


Figure 5.7: Subjective MUSHRA quality ratings (averaged over sentences) for speech in the *high predicted benefit* case.

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.

5.7.2.1 *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 5.1. 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 ($p > 0.999$).

5.7.2.2 *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 5.2. 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.

5.7.3 *Prediction of subjective quality ratings*

This section investigates how well the objective metrics described in Sect. 5.4 predict the subjective ratings described in Sect. 5.7.1. The prediction performance was

Table 5.1: 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 5.2: 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

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 ρ_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_3) to account for systematic variation in subjective ratings. For computing RMSE values, the predictions were scaled to the MUSHRA range to facilitate comparisons between metrics². For PEMO-Q, an output 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 Fig. 5.8. 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^a, 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.

²It 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.

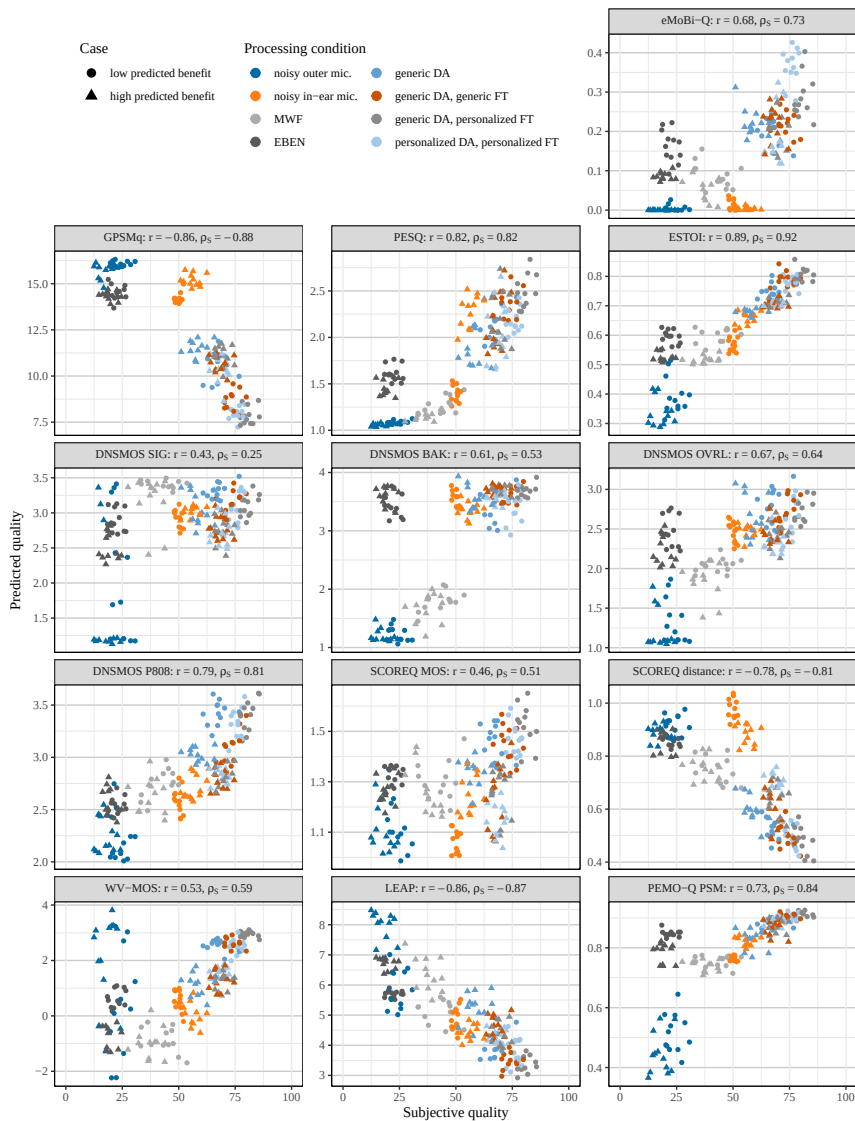


Figure 5.8: Scatterplots comparing median subjective quality ratings (abscissae) and predicted objective quality (ordinates) for noisy and reconstructed own voice signals.

Table 5.3: Correlation coefficients and RMSE values for the considered metrics.

Metric	r	ρ_S	RMSE	RMSE ₃
eMoBi-Q	0.68	0.73	40.8	15.2
GPSM ^a	-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 distance	-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

Correlation coefficients as well as root mean squared error (RMSE) values are reported in Table 5.3. In terms of linear correlation, ESTOI predictions achieved the highest (absolute) correlation coefficient, followed by LEAP, GPSM^a, 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 Fig. 5.8, many of the metrics do not predict close to the upper end of their respective 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 RMSE₃ values after fitting 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^a, PESQ, and SCOREQ distance.

5.8 Discussion

5.8.1 *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 [17] revealed low correlation of PESQ and SI-SDR (a technical metric) with subjective ratings, but higher correlations were observed for STOI and Noresqa-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 [19] and [17], 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.

5.8.2 *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.

5.8.3 *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 Fig. 5.8

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]³, 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 and not for quality predictions. 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. Comparing the results of eMoBi-Q and GPSM^q, we observe that GPSM^q achieves higher correlations and lower errors, which is likely due to the ability of GPSM^q to account for non-linear moderate or heavy distortions.

5.9 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 band-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.

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

Author declarations

Conflicts of interest

The authors declare no conflict of interest.

Informed consent

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

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/.

Acknowledgments

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. 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.

References

- [1] R. E. Bouserhal, T. H. Falk, and J. Voix, “Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones,” in *Proc. Meetings on Acoustics (ICA)*, vol. 19, Montreal, QC, Canada, 2013.
- [2] S. Nordholm, A. Davis, P. C. Yong, and H. H. Dam, “Assistive listening headsets for high noise environments: Protection and communication,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5753–5757.

- [3] R. E. Bouserhal, A. Bernier, and J. Voix, “An In-Ear Speech Database in Varying Conditions of the Audio-Phonation Loop,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 1069–1077, Feb. 2019.
- [4] B. Gårdbæk and P. Kidmose, “On the origin of cardiovascular sounds recorded from the ear,” *IEEE Trans. on Biomedical Engineering*, vol. 72, no. 1, pp. 210–216, 2025.
- [5] M. Ø. Hansen, “Occlusion effects Part I and II,” PhD thesis, Department of Acoustic Technology, Technical University of Denmark, 1998.
- [6] T. Zurbrügg, A. Stirnemann, M. Kuster, and H. Lissek, “Investigations on the physical factors influencing the ear canal occlusion effect caused by hearing aids,” *Acta Acustica united with Acustica*, vol. 100, no. 3, pp. 527–536, May 2014.
- [7] S. Stenfelt and S. Reinfeldt, “A model of the occlusion effect with bone-conducted stimulation,” *International Journal of Audiology*, vol. 46, no. 10, pp. 595–608, Jan. 2007.
- [8] S. Vogl and M. Blau, “Individualized prediction of the sound pressure at the eardrum for an earpiece with integrated receivers and microphones,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 917–930, Feb. 2019.
- [9] S. Reinfeldt, P. Östli, B. Håkansson, and S. Stenfelt, “Hearing one’s own voice during phoneme vocalization - Transmission by air and bone conduction,” *J. Acoust. Soc. Am.*, vol. 128, no. 2, pp. 751–762, Aug. 2010.
- [10] H. Saint-Gaudens, H. Nélisse, F. Sgard, and O. Doutres, “Towards a practical methodology for assessment of the objective occlusion effect induced by earplugs,” *J. Acoust. Soc. Am.*, vol. 151, no. 6, pp. 4086–4100, Jun. 2022.
- [11] J. Richard, V. Zimpfer, C. Blondé-Weinmann, and S. Roth, “Change in transfer function between air and bone conduction microphones due to mouth opening variation,” *Applied Acoustics*, vol. 228, p. 110 293, Jan. 2025.
- [12] F. Denk and B. Kollmeier, “The Hearpiece database of individual transfer functions of an in-the-ear earpiece for hearing device research,” *Acta Acustica*, vol. 5, no. 2, 2021.
- [13] K. Kondo, T. Fujita, and K. Nakagawa, “On equalization of bone conducted speech for improved speech quality,” in *Proc. International Symposium on Signal Processing and Information Technology*, Vancouver, BC, Canada, Aug. 2006, pp. 426–431.
- [14] H. S. Shin, T. Fingscheidt, and H.-G. Kang, “A priori SNR estimation using air- and bone-conduction microphones,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2015–2025, Nov. 2015.
- [15] R. E. Bouserhal, T. H. Falk, and J. Voix, “In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension,” *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 1321–1331, Mar. 2017.

- [16] H. Wang, X. Zhang, and D. Wang, “Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 3134–3143, 2022.
- [17] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, “Configurable EBEN: Extreme Bandwidth Extension Network to Enhance Body-conducted Speech Capture,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 3499–3512, 2023.
- [18] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Multi-Microphone Noise Data Augmentation for DNN-based Own Voice Reconstruction for Hearables in Noisy Environments,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 416–420. Accessed: Dec. 18, 2023.
- [19] C. Li, F. Yang, and J. Yang, “A two-stage approach to quality restoration of bone-conducted speech,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 818–829, 2024.
- [20] C. Li, F. Yang, and J. Yang, “Restoration of bone-conducted speech with U-Net-like model and energy distance loss,” *IEEE Signal Processing Letters*, vol. 31, pp. 166–170, 2024.
- [21] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, “Speaker Adaptation For Enhancement Of Bone-Conducted Speech,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 10 456–10 460. Accessed: Mar. 26, 2024.
- [22] Y. Sui, M. Zhao, J. Xia, X. Jiang, and 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 *Proc. ACM Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, New York, USA, Nov. 2024.
- [23] J. Richard, V. Zimpfer, and S. Roth, “Comparison of objective and subjective methods for evaluating speech quality and intelligibility recorded through bone conduction and in-ear microphones,” *Applied Acoustics*, vol. 211, Aug. 2023. Accessed: Oct. 10, 2023.
- [24] K. Tesch and T. Gerkmann, “Insights into deep non-linear filters for improved multi-channel speech enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023.
- [25] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Speech-dependent Data Augmentation for Own Voice Reconstruction with Hearable Microphones in Noisy Environments,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, no. 32, 2025.
- [26] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Low-complexity own voice reconstruction for hearables with an in-ear microphone,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025.

- [27] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Modeling of speech-dependent own voice transfer characteristics for hearables with an in-ear microphone,” *Acta Acust.*, vol. 8, no. 28, 2024. [Online]. Available: <https://doi.org/10.1051/aacus/2024032>.
- [28] F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J.-H. Bach, J. Wellmann, and B. Kollmeier, “A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research,” in *Proc. AES International Conference on Headphone Technology*, San Francisco, USA, Aug. 2019.
- [29] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common Voice: A massively-multilingual speech corpus,” in *Proc. 12th Language Resources and Evaluation Conference*, Marseille, France, May 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, and R. Aichner, “ICASSP 2023 Deep Noise Suppression Challenge,” *IEEE Open Journal of Signal Processing*, vol. 5, pp. 725–737, 2024. Accessed: Mar. 27, 2024.
- [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,” *International Telecommunications Union*, Feb. 2001, Geneva, Switzerland.
- [32] International Telecommunications Union (ITU), “ITU-T P.863, Perceptual objective listening quality prediction (POLQA),” *International Telecommunications Union*, Mar. 2018, Geneva, Switzerland.
- [33] J. F. Santos, R. Bouserhal, J. Voix, and T. H. Falk, “Objective speech quality estimation of in-ear microphone speech,” in *Proc. 5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)*, Berlin, Germany, Aug. 2016, pp. 69–73.
- [34] P. Bauer, C. Guillaumé, W. Tirry, and T. Fingscheidt, “On speech quality assessment of artificial bandwidth extension,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, Jul. 2014, pp. 6082–6086.
- [35] J. Jensen and C. H. Taal, “An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
- [36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [37] T. Biberger, J.-H. Fleßner, R. Huber, and S. D. Ewert, “An objective audio quality measure based on power and envelope power cues,” *J. Audio Eng. Soc.*, vol. 66, no. 7/8, pp. 578–593, Jul. 2018.

- [38] B. Eurich, S. D. Ewert, M. Dietz, and T. Biberger, “A computationally efficient model for combined assessment of monaural and binaural audio quality,” *J. Audio Eng. Soc.*, vol. 72, no. 9, pp. 536–551, Sep. 2024.
- [39] J.-H. Fleßner, T. Biberger, and S. D. Ewert, “Subjective and objective assessment of monaural and binaural aspects of audio quality,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1112–1125, 2019.
- [40] T. Dau, D. Püschel, and A. Kohlrausch, “A quantitative model of the “effective” signal processing in the auditory system. I. Model structure,” *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, 1996.
- [41] R. Huber and B. Kollmeier, “PEMO-Q — A new method for objective audio quality assessment using a model of auditory perception,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.
- [42] A. Ragano, J. Skoglund, and 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, and R. Cutler, “DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 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,” *International Telecommunications Union*, 2003, Geneva, Switzerland.
- [45] International Telecommunications Union (ITU), “ITU-T P.808, Subjective evaluation of speech quality with a crowdsourcing approach,” *International Telecommunications Union*, 2018, Geneva, Switzerland.
- [46] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, “HIFI++: A unified framework for bandwidth extension and speech enhancement,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 2023.
- [47] R. Huber, M. Krüger, and B. T. Meyer, “Single-ended prediction of listening effort using deep neural networks,” *Hearing Research*, vol. 359, pp. 40–49, 2018.
- [48] R. Huber, A. Pusch, N. Moritz, J. Rennies, H. Schepker, and B. T. Meyer, “Objective assessment of a speech enhancement scheme with an automatic speech recognition-based system,” in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 86–90.
- [49] M. Krueger, M. Schulte, T. Brand, and I. Holube, “Development of an adaptive scaling method for subjective listening effort,” *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4680–4693, 2017.

- [50] International Telecommunications Union (ITU), “ITU-R BS.1534-3, method for the subjective assessment of intermediate sound quality (MUSHRA),” *International Telecommunications Union*, Oct. 2015, Geneva, Switzerland.
- [51] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, “webMUSHRA - A comprehensive framework for web-based listening tests,” *Journal of Open Research Software*, vol. 6, no. 1, Feb. 2018.
- [52] S. Doclo and M. Moonen, “GSVD-based optimal filtering for single and multi-microphone speech enhancement,” *IEEE Trans. on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.
- [53] M. Souden, J. Chen, J. Benesty, and S. Affes, “Gaussian model-based multichannel speech presence probability,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, 2009.
- [54] S. Bagheri and D. Giacobello, “Exploiting multi-channel speech presence probability in parametric multi-channel wiener filter,” in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 101–105.
- [55] J. Rennies, M. Ohlenbusch, A. Volgenandt, T. Spitz, H. Baumgartner, C. Rollwage, V. Uslar, and V. Weber, “Analyse und algorithmische Optimierung von Geräuschkulissen und Sprachkommunikation im OP-Saal,” in *Proc. German Annual Conference on Acoustics (DAGA)*, Hamburg, Germany, Mar. 2023, pp. 646–649.
- [56] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2025. [Online]. Available: <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, and J. Berger, “Subjective and objective assessment of full bandwidth speech quality,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 440–449, 2019.
- [58] L. He, H. Hou, S. Shi, X. Shuai, and Z. Yan, “Towards bone-conducted vibration speech enhancement on head-mounted wearables,” in *Proc. Annual International Conference on Mobile Systems, Applications and Services*, New York, USA, Jun. 2023, pp. 14–27.
- [59] A. Kuznetsova, A. Sivaraman, and M. Kim, “The potential of neural speech synthesis-based data augmentation for personalized speech enhancement,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 2023.
- [60] J.-S. Bae, A. Kuznetsova, D. Manocha, J. Hershey, T. Kristjansson, and M. Kim, “Generative data augmentation challenge: Zero-shot speech synthesis for personalized speech enhancement,” in *Proc. International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW): Generative Data Augmentation for Real-World Signal Processing Applications (GenDA 2025)*, Hyderabad, India, Apr. 2025.

- [61] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, “Neural target speech extraction: An overview,” *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.
- [62] J. Hauret, M. Olivier, T. Joubaud, C. Langrenne, S. Poirée, V. Zimpfer, and É. Bavu, “Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors,” *Speech Communication*, Apr. 2025.
- [63] H. Pulakka, V. Myllylä, A. Rämö, and P. Alku, “Speech quality evaluation of artificial bandwidth extension: Comparing subjective judgments and instrumental predictions,” in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 2583–2587.
- [64] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, “HIFI++: A unified framework for bandwidth extension and speech enhancement,” *arXiv:2203.13086*, Dec. 2023.
- [65] J. Rennies, M. Berdau, R. Huber, H. Baumgartner, S. Weihe, and T. Brand, “Real-time assessment of listening effort using non-intrusive binaural prediction models,” in *Proc. DAS/DAGA 2025*, Copenhagen, Denmark, Mar. 2025, pp. 18–21.
- [66] M. Ohlenbusch, C. Rollwage, and S. Doclo, *German own voice recordings with hearable microphones*, Zenodo, Mar. 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.10844599>.
- [67] M. Ohlenbusch, C. Rollwage, and S. Doclo, *Transfer function measurements for simulating environmental noise at hearable microphones*, Zenodo, May 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.11196867>.
- [68] M. Ohlenbusch, C. Rollwage, S. Doclo, and J. Rennies, *Subjective ratings and objective metric predictions of generic and personalized own voice reconstruction systems*, Zenodo, Apr. 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.15248719>.

PAS-SE: PERSONALIZED AUXILIARY-SENSOR SPEECH ENHANCEMENT FOR VOICE PICKUP IN HEARABLES

This chapter is identical in content to the publication: M. Ohlenbusch, M. Kegler, and M. Stamenovic, “PAS-SE: Personalized auxiliary-sensor speech enhancement for voice pickup in hearables,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2026, pp. 18 942–18 946. DOI: [10.1109/ICASSP55912.2026.11460554](https://doi.org/10.1109/ICASSP55912.2026.11460554).

Authors	Author's contribution							
	A	B	C	D	E	F	G	H
Mattes Ohlenbusch	✗	✗	✗	✗	✗	✗	✗	✗
Mikolaj Kegler	✗		✗	✗	✗	✗	✗	✗
Marko Stamenovic	✗		✗	✗		✗	✗	✗

A - Substantial contributions to the conception or design of the work

B - Acquisition of the data

C - Analysis of the data

D - Interpretation of the data

E - Drafting the work

F - Revising the work critically

G - Final approval of the version to be published

H - Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Abstract

Speech enhancement for voice pickup in hearables aims to improve the user’s voice by suppressing noise and interfering talkers, while maintaining own-voice quality. For single-channel methods, it is particularly challenging to distinguish the target from interfering talkers without additional context. In this paper, we compare two strategies to resolve this ambiguity: personalized speech enhancement (PSE), which uses enrollment utterances to represent the target, and auxiliary-sensor speech enhancement (AS-SE), which uses in-ear microphones as additional input. We evaluate the strategies on two public datasets, employing different auxiliary sensor arrays, to investigate their cross-dataset generalization. We propose training-time augmentations to facilitate cross-dataset generalization of AS-SE systems. We also show that combining PSE and AS-SE (PAS-SE) provides complementary performance benefits, especially when enrollment speech is recorded with the in-ear microphone. We further demonstrate that PAS-SE personalized with noisy in-ear enrollments maintains performance benefits over the AS-SE system.

6.1 Introduction

Hearable devices with one or more microphones can be used to capture the user’s own voice (i.e., target speech), but due to environmental noise and interfering talkers, the captured speech signal needs to be enhanced to facilitate communication. Although single-channel speech enhancement (SE) approaches based on deep neural networks can reduce environmental noise [1–4], they tend to struggle with removing interfering talkers from the mixture while preserving the user’s voice. Modern hearables often include microphones placed inside or near the user’s occluded ear canal, typically employed in active noise reduction systems [5]. These in-ear microphones are acoustically shielded from environmental noise and interfering talkers by the device. At the same time, the user’s voice is also picked up by the in-ear microphone, predominantly through body conduction [6, 7]. These two effects lead to a substantial benefit in terms of signal-to-noise ratio (SNR) of the user’s voice at the in-ear microphone. However, this signal cannot be directly used for voice communication due to time- and user-varying band-limitation, nonlinearities and distortions introduced by body conduction [6, 7], and undesired additive body-produced noises [8]. Other auxiliary body-conduction sensors, such as accelerometers, have similar properties and trade-offs as in-ear microphones [9].

Due to their benefits, auxiliary sensors have been employed for own-voice speech enhancement [9–13]. The use of an auxiliary sensor as an additional input can substantially improve performance, especially in challenging SNRs [14] and scenarios with interfering talkers [9]. Notably, in [9] it was found that using a body-conduction sensor as additional input facilitated interferer suppression when tested with simulated data, assuming no noise and interferer transmission to the auxiliary sensor. We broadly refer to this approach, where the system uses both an outer microphone and an auxiliary sensor as input, as auxiliary-sensor speech enhancement (AS-SE). Another way to extract speech from a target talker in a single-channel system is to condition the system with a feature vector obtained from enrollment utterances of the target talker [15–19], commonly referred to as target speaker extraction (TSE). In this work, we consider a special case of TSE, personalized speech enhancement (PSE), where the target talker is always the device user. Using the enrollment utterance of the user, the PSE system can resolve the ambiguity between target and interfering talker.

It should be noted that the methods mentioned above each come with their own unique trade-offs. PSE typically requires enrollment speech, which implies an additional setup procedure for the user prior to using the system. In contrast, AS-SE systems can be readily used by any user without an additional setup procedure. However, the resulting system may not generalize across different devices due to unique array properties or acoustic design of the hearable device.

A systematic evaluation or integration of PSE and AS-SE methods has yet to be carried out. In this paper, we:

1. Benchmark PSE and AS-SE performance by systematically evaluating their denoising capabilities, suppression of interfering talkers, and generalization across datasets.

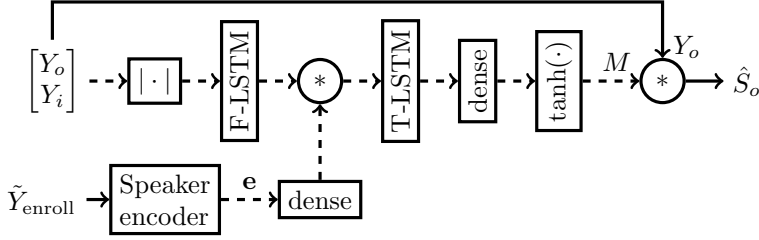


Figure 6.1: PAS-SE system architecture based on FT-JNF [20]. The system is personalized using multiplicative conditioning with a feature vector $\mathbf{e}$ obtained from an enrollment utterance $\tilde{Y}_{\text{enroll}}$.

2. Explore and ablate different training-time augmentation configurations for training AS-SE systems with limited pre-recorded in-ear signals to facilitate interfering talker suppression and cross-dataset generalization.
3. Personalize AS-SE through enrollment-based conditioning, investigate the benefits of auxiliary-sensor enrollment to develop personalized AS-SE (PAS-SE) systems, and analyze their robustness to noise in enrollment utterances.

Experimental evaluations carried out on two openly available in-ear microphone datasets demonstrate that interfering speech can be considerably reduced with either PSE or AS-SE, while a combination of both, which we term personalized auxiliary-sensor speech enhancement (PAS-SE), leads to further improvements. In particular, PAS-SE personalized with in-ear enrollments leads to the best results, generalizing both within and across datasets tested.

6.2 Signal model

We consider a hearable device equipped with an outer microphone and an in-ear microphone. The signals are denoted by subscripts o for the outer microphone (OM) and i for the in-ear microphone (IM). In the short-time Fourier transform (STFT) domain, $S_o(k, l)$ and $S_i(k, l)$ denote the own voice signals of the user (i.e., target speech) at each microphone, where k and l denote the frequency and the frame index. The outer and in-ear microphone signals are given by

$$Y_{\{o,i\}}(k, l) = S_{\{o,i\}}(k, l) + N_{\{o,i\}}(k, l) + V_{\{o,i\}}(k, l), \quad (6.1)$$

respectively, where the noise components are denoted by $N_o(k, l)$ and $N_i(k, l)$, and the interfering talker components are denoted by $V_o(k, l)$ and $V_i(k, l)$.

6.3 System architecture

In this paper, we explore various SE, PSE, AS-SE, and PAS-SE systems by modifying the FT-JNF architecture proposed in [20]. This architecture has previously been applied to the AS-SE task in [14, 21]. Fig. 6.1 shows the personalized FT-JNF

variant used in this paper, which takes the magnitude of the noisy STFT coefficients of one or more microphones as input. The architecture consists of a unidirectional long short-term memory (LSTM) layer operating across the frequency dimension of the input (F-LSTM) with 512 hidden units, a causal unidirectional LSTM layer operating across the time dimension (T-LSTM) with 128 hidden units, a linear layer, and a tanh activation function. The output is a magnitude mask $M(k, l)$, which is applied to the noisy outer microphone signal $Y_o(k, l)$, i.e. $\hat{S}_o(k, l) = M(k, l) \cdot Y_o(k, l)$. The SE system (with one input) has 1.384M parameters, and the AS-SE system has 1.386M parameters.

We add a conditioning mechanism to the architecture to investigate personalization for our PSE and PAS-SE systems by introducing a speaker encoder branch based on the time-domain SpeakerBeam architecture [15] using the authors' implementation in the Asteroid toolkit [22]. An utterance $\tilde{Y}_{\text{enroll}}(k, l)$ from the same user, excluding the input utterance (either recorded at the outer or in-ear microphone), is used as input to the speaker encoder branch. The branch consists of a learnable filterbank encoder, which is followed by a 1-D convolutional block and temporal averaging to obtain a 128-dimensional speaker embedding vector $\mathbf{e}$. A dense layer is then used to match the dimension of the output of the F-LSTM layer of the main enhancement network. The output of the F-LSTM layer is multiplied with the output of the dense layer from the speaker encoder branch. The speaker encoder consists of 1.810M parameters.

6.4 Evaluation details

6.4.1 Datasets

The Vibravox Dataset [23] was used for training and evaluation. It contains approximately 33h of clean speech signals recorded from 198 talkers, for which we used the official split into training, validation, and test data of the *speech_clean* subset. Recorded signals from the rigid earbud with an in-ear microphone [24] (*rigid in-ear*) were used for training. The close-talk microphone (a headset boom microphone) was selected in place of an outer microphone. For environmental noise, we use the *speechless_noisy* subset of Vibravox.

The Oldenburg dataset [7] was used for the cross-dataset evaluation of the proposed systems (and training dataset-specific baselines). The dataset consists of 306 utterances by 12, 2, and 4 different talkers across training, validation, and test, respectively. The dataset was recorded using the same type of rigid earbud with an in-ear microphone [24] as Vibravox. Unlike Vibravox, which uses the close-talk mic, a microphone at the outer face of the device (most common in commodity hearables) is used as outer microphone. This induces a large difference in the auxiliary sensor array between the in-ear and outer microphones compared to the Vibravox setup. For environmental noise, noise signals from the fifth DNS challenge [25] were convolved with the individual impulse responses of the device user to obtain spatialized multi-channel noise signals. The impulse responses consisted of measurements from 8 evenly spaced loudspeakers in a circle around the user [14]. For interfering

Table 6.1: Configurations of additive noise and interferer approximations at the outer and in-ear microphones for training AS-SE and PAS-SE systems. All training configurations include own voice recorded at both in-ear and outer microphones $S_{\{o,i\}}$.

	Outer microphone		In-ear microphone	
	Noise N_o	Interferer V_o	Noise N_i	Interferer V_i
A [9]	✓	✓	✗	✗
B [14]	✓	✗	✓	✗
C	✓	✓	✓	✗
D	✓	✓	✓	$a \cdot V_o$

talkers, the same procedure is followed replacing noise signals with talkers from the Oldenburg training set other than the target talker.

The same procedure for mixing own voice, noise, and interferers is carried out for both datasets. For each utterance during training, there is a 75% probability of environmental noise being added. If added, own voice and environmental noise are mixed to an SNR randomly selected from a uniform distribution between $[-10, 10]$ dB (defined at the outer microphone). The corresponding scaling factor is applied to the in-ear microphone signals as well, to realistically preserve SNR differences between the noisy outer and acoustically shielded in-ear microphone. The same procedure is applied to interferers, with an independent 75% probability. This ensures the system is trained on a combination of user’s clean speech, noisy speech, noisy speech with interferers, and speech with only interferers. Since Vibravox does not contain an isolated interfering speaker partition, we randomly select from the set of speakers in the training partition disjoint from the target speaker as the interferer at each step.

6.4.2 In-ear noise and interferer training configurations

To facilitate training of AS-SE and PAS-SE systems without recorded interferer signals (as in Vibravox), we explore various configurations of incorporating in-ear signal noise and interferer components during training as shown in Table 6.1. The components in (6.1) are either included using recorded signals (✓) or not (✗) during training, or the in-ear microphone component is approximated by the corresponding outer microphone component attenuated using a random factor $a \in [0.001, 1]$ from a uniform distribution.

In configuration (A), we consider training an AS-SE system by adding noise N_o and interfering speech V_o only to the outer microphone, but not to the auxiliary sensor signal as in [9]. Configuration (B) consists of adding noise $N_{\{o,i\}}$ to the outer and in-ear microphone signals, but no interfering talkers to either sensor as in [14]. In configuration (C), we consider training with both noise components

$N_{\{o,i\}}$ and adding interferers V_o only to the outer microphone signal. Finally, in configuration (D), we use all components during training but approximate the in-ear interferer component by an attenuated version of the outer microphone component as $V_i \approx a \cdot V_o$.

6.4.3 Experimental setup

The experiments are conducted at a sampling rate of 16 kHz, using an STFT framework with a frame length of 32 ms and a frame shift of 16 ms, where a square-root Hann window is used both in analysis and synthesis. Training is carried out with a batch size of 8 and an example length of 3 seconds (randomly selected clip). Mean-variance-normalization based on clean speech training subset statistics is applied to each input channel independently. For personalized systems, a single disjoint utterance $\tilde{S}$ from the same talker was randomly selected as the clean enrollment speech $\tilde{Y}_{\text{enroll}} = \tilde{S}_{\text{enroll}\{o,i\}}$ (OM or IM enrollment) for each training example.

The loss consists of the combined L_1 difference between clean target speech at the outer microphone S_o and the estimated speech signal $\hat{S}_o$ in both time and STFT magnitude domains [26]. Gradient clipping is used if the total gradient L_2 -norm is greater than 10. The ADAM optimizer [27] is used with an initial learning rate of 0.001, which is halved every five epochs until training for a total of 50 epochs, after which the system from the epoch with the best validation loss is selected. We use the time-domain SpeakerBeam [15] architecture (4.985 M parameters without the speaker encoder) as a baseline system and train it with the same setups described above.

6.5 Results

For evaluation, we use the scale-invariant signal-to-distortion ratio (SI-SDR) [28], wideband PESQ [29], and ESTOI [30] between the target and estimated clean speech signal at the outer microphone¹. Results are averaged across the entire test set. For evaluation, target speech is either mixed with noise (**N**), interfering voices (**V**) or both (**N+V**). To evaluate robustness against noisy enrollment speech in Section 6.5.3, clean enrollment utterances are mixed with noise $\tilde{N}$, so $\tilde{Y}_{\text{enroll}} = \tilde{S}_{\text{enroll}\{o,i\}} + \tilde{N}_{\{o,i\}}$, and the performance of the system is evaluated only adding interfering speech (**V**) to the target speech.

6.5.1 In-domain evaluation

Table 6.2 shows the results for SE and PSE systems evaluated on the Vibravox dataset. All systems meaningfully improve speech quality in terms of objective metrics when only noise is present. For the SE system, in the cases of interferer or interferer and noise, only slight improvements over the noisy signals are achieved, indicating the system does not have enough information to suppress interferers in

¹ Audio examples & paper resources: <https://bose.github.io/passe/>.

Table 6.2: Vibravox evaluation results for target speech mixed with noise (**N**), with an interferer (**V**), or with both (**N+V**). Personalization based on outer and in-ear microphone enrollment utterances is denoted by OM and IM, respectively. SB: SpeakerBeam [15].

System	Enrol.	<i>SI-SDR (dB)</i>			<i>PESQ</i>			<i>ESTOI</i>		
		N	V	N+V	N	V	N+V	N	V	N+V
Noisy	-	0.10	-0.04	-4.50	1.21	1.24	1.11	0.53	0.54	0.35
SE	-	7.64	0.55	-1.76	1.67	1.31	1.20	0.62	0.53	0.38
SB [15]	OM	7.87	5.21	1.83	1.49	1.39	1.19	0.60	0.61	0.41
	IM	7.64	4.28	0.82	1.47	1.35	1.18	0.59	0.59	0.39
PSE	OM	8.72	5.78	2.65	1.74	1.56	1.31	0.65	0.64	0.47
	IM	8.26	4.81	1.77	1.72	1.52	1.29	0.64	0.63	0.45

these scenarios. On the other hand, enhancement using the PSE or SpeakerBeam systems yields meaningful interferer reduction. When comparing the outer with the in-ear microphone for enrollment, both systems are observed to benefit more from personalization with the outer microphone. Despite being smaller, the FT-JNF PSE systems achieve slightly higher scores than SpeakerBeam in all scenarios.

Table 6.3 shows the results for noise reduction by AS-SE and PAS-SE systems trained with different configurations of noise and interferer in-ear components evaluated on the Vibravox dataset. When no in-ear noise or interferer components are used (A), performance is poor, but including in-ear noise in training (B, C, D) yields large improvements, highlighting the importance of modeling the noise leakage to the in-ear microphone during training. As expected for noise reduction, different configurations of interferer in-ear components do not lead to substantial performance differences between configurations (B), (C), and (D). Similarly, added personalization (PAS-SE) does not yield improvements.

6.5.2 Cross-dataset generalization

Table 6.4 shows the cross-dataset generalization (out-of-domain) results of various Vibravox-trained systems evaluated on the Oldenburg dataset. Again, the non-personalized SE system is able to reduce noise, but achieves little improvement in terms of reducing interfering talkers. The PSE baseline SpeakerBeam trained on Vibravox fails to enhance speech on the Oldenburg dataset, exhibiting SI-SDR scores below that of noisy signals, while FT-JNF-based PSE trained in the same way achieves similar scores on both Oldenburg and Vibravox test sets. This may be due to the fact that SpeakerBeam operates in the time domain, employing learnable filterbanks, which may be more prone to dataset-specific biases, while FT-JNF uses magnitude STFT features.

Table 6.3: Vibravox evaluation results of FT-JNF-based AS-SE and PAS-SE systems optimized with different training configurations. Only noise suppression is evaluated, since the Vibravox dataset does not contain in-ear microphone signals of isolated interfering voices.

System	Enrol.	Train.	Configuration				<i>SI-SDR</i> (dB)	<i>PESQ</i>	<i>ESTOI</i>
			N_o	V_o	N_i	V_i	N	N	N
Noisy	-		N/A				0.10	1.21	0.53
AS-SE	-	A [9]	✓	✓	✗	✗	-0.92	1.25	0.50
	-	B [14]	✓	✗	✓	✗	11.23	2.19	0.75
	-	C	✓	✓	✓	✗	10.42	2.06	0.73
	-	D	✓	✓	✓	$a \cdot V_o$	9.98	2.02	0.72
	OM IM	C	✓	✓	✓	✗	10.68 10.88	2.09 2.14	0.73 0.74
PAS-SE	OM IM	D	✓	✓	✓	$a \cdot V_o$	10.29 10.29	2.07 2.06	0.72 0.72

For the AS-SE system trained considering noise and interferer only at the outer microphone (A), performance is unsurprisingly poor, due to the fact that in-ear noise or interferers were not present during training. In configuration (B), where the system is only trained with environmental noise and without interferers, performance for noise reduction improves, whereas interferer reduction does not. When an interfering talker is added only at the outer microphone along with noise at both sensors (C), both noise and interferer reduction performance improves, indicating that interferer modeling at the in-ear microphone may not be strictly necessary to enable interferer rejection. However, the best performance is achieved when in-ear interferers are approximated and added to recorded in-ear noise (D).

In the PAS-SE systems, we observe systematic performance improvements compared to their AS-SE counterparts, especially on but not limited to interferers. For a PAS-SE system trained with configuration (C), interferer suppression improves by conditioning with either outer or in-ear microphone enrollments, indicating the conditioning may compensate for the absence of interfering voice leakage to the in-ear microphone during training. For configuration (D), further improvements are observed compared to (C) in *SI-SDR* for interferers and noise and interferer, while other metrics remain similar. Interestingly, for systems both trained and evaluated on the Oldenburg data (OL), personalization yields smaller improvements, likely due to only 12 speakers being available in the training subset.

Overall, the performance trends indicate that the AS-SE systems require the use of some level of interferer signals during training (C, D) to generalize to an unseen auxiliary-sensor array. Cross-domain performance improves more when combined

Table 6.4: Oldenburg evaluation results for target speech mixed with noise ($\mathbf{N}$), with an interferer ($\mathbf{V}$), and with both interferer and noise ($\mathbf{N}+\mathbf{V}$). All models trained on Vibravox data, except those denoted by OL where the system is trained on the in-domain Oldenburg training dataset.

System	Enrol.	Train.	Configuration				$SI\text{-}SDR$ (dB)			$PESQ$			$ESTOI$		
			N_o	V_o	N_i	V_i	$\mathbf{N}$	$\mathbf{V}$	$\mathbf{N}+\mathbf{V}$	$\mathbf{N}$	$\mathbf{V}$	$\mathbf{N}+\mathbf{V}$	$\mathbf{N}$	$\mathbf{V}$	$\mathbf{N}+\mathbf{V}$
Noisy	-	-					0.13	0.04	-4.52	1.28	1.51	1.18	0.40	0.55	0.30
SE	-	-					8.04	2.23	-0.06	1.60	1.58	1.31	0.46	0.57	0.34
SpeakerBeam [15]	OM	-	N/A				-1.42	-7.02	-8.63	1.19	1.24	1.13	0.28	0.39	0.19
	IM	-					-1.64	-5.21	-7.44	1.16	1.20	1.11	0.28	0.38	0.19
PSE	OM	-					8.29	4.67	2.01	1.63	1.70	1.36	0.47	0.58	0.36
	IM	-					8.43	4.45	2.00	1.61	1.67	1.35	0.47	0.58	0.37
AS-SE	-	A [9]	✓	✓	✗	✗	3.35	3.52	-1.72	1.51	1.78	1.33	0.49	0.63	0.41
	-	B [14]	✓	✗	✓	✗	9.85	2.62	2.60	1.87	1.67	1.50	0.56	0.60	0.46
	-	C	✓	✓	✓	✗	10.09	5.15	3.59	1.88	1.90	1.57	0.56	0.64	0.48
	-	D	✓	✓	✓	$a \cdot V_o$	9.63	7.20	4.97	1.86	1.93	1.60	0.55	0.64	0.47
PAS-SE	OM	C	✓	✓	✓	✗	10.57	6.38	4.87	1.92	1.95	1.62	0.56	0.65	0.48
	IM		✓	✓	✓		10.70	7.40	5.35	1.92	2.01	1.64	0.57	0.66	0.49
	OM	D	✓	✓	✓	$a \cdot V_o$	9.24	7.17	5.31	1.85	1.89	1.61	0.54	0.64	0.48
	IM		✓	✓	✓		10.30	8.34	5.85	1.88	1.98	1.62	0.55	0.65	0.48
AS-SE	-	OL					7.47	7.27	4.57	1.88	2.08	1.72	0.51	0.63	0.46
PAS-SE	OM	OL					7.45	7.49	4.64	1.97	2.19	1.82	0.53	0.64	0.49
	IM						7.63	7.57	4.85	2.00	2.23	1.84	0.53	0.64	0.49

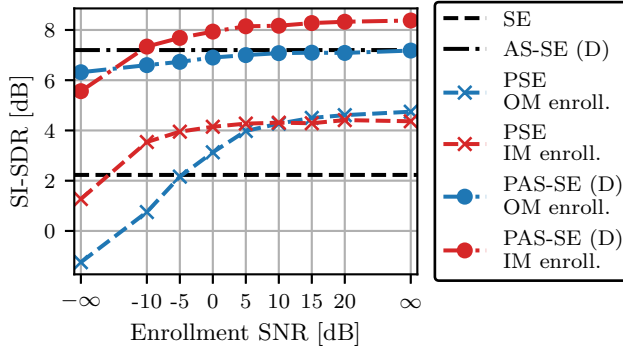


Figure 6.2: Cross-dataset interferer reduction performance ($\mathbf{V}$) achieved by SE, PSE, AS-SE, and PAS-SE systems at different enrollment utterance SNRs ($-\infty$: only noise, no speech, ∞ : clean speech).

with enrollment-based personalization (PAS-SE), nearing or even outperforming the corresponding matched system trained on in-domain data (OL).

6.5.3 Personalization with noisy enrollment signals

The performance of the proposed systems personalized with either in-ear or outer microphone enrollment utterances was evaluated with noisy enrollment signals (Fig. 6.2). Similar to Table 6.4, the systems were trained on Vibravox and evaluated on Oldenburg datasets (out-of-domain). For PAS-SE and AS-SE, we use the model trained with configuration (D), which yielded the overall best results.

For PSE systems, conditioning with an in-ear microphone is more robust to noise in the enrollment utterance, likely due to the in-ear microphone being acoustically shielded from noise. Using the in-ear microphone yields improvements over the non-personalized single-channel system even at -10 dB enrollment SNR, whereas using the outer microphone for conditioning provides no benefits below 0 dB. Similarly, the PAS-SE systems conditioned with the in-ear enrollment speech achieve better performance than the AS-SE systems for enrollment SNRs higher than -10 dB, while PAS-SE systems conditioned with an outer microphone do not benefit from personalization at all. While not considered here, training with enrollment augmentation as in [31] may provide further benefits.

6.6 Conclusion

In this paper, we have systematically evaluated PSE and AS-SE in terms of denoising, interferer suppression, and cross-dataset generalization. Experimental results demonstrate that the proposed training-time augmentations can yield substantial cross-dataset generalization benefits for AS-SE systems, nearing the performance

of systems trained fully in-domain. Moreover, the proposed PAS-SE systems with enrollment-based personalization generalize across datasets, outperforming dataset-specific baselines, and retain their benefits with in-ear enrollments even when the enrollment contains noise. We hope that the proposed methodology can facilitate future research in the direction of fully device-agnostic AS-SE systems.

References

- [1] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, “Towards efficient models for real-time deep noise suppression,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 656–660.
- [2] X. Rong, T. Sun, X. Zhang, Y. Hu, C. Zhu, and J. Lu, “GTCRN: A speech enhancement model requiring ultralow computational resources,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 971–975.
- [3] I. Fedorov, M. Stamenovic, C. Jensen, L.-C. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, “TinyLSTMs: Efficient neural speech enhancement for hearing aids,” in *Proc. Interspeech*, 2020, pp. 4054–4058.
- [4] R. D. Nathoo, M. Kegler, and M. Stamenovic, “Two-step knowledge distillation for tiny speech enhancement,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 141–10 145.
- [5] C.-Y. Chang, A. Siswanto, C.-Y. Ho, T.-K. Yeh, Y.-R. Chen, and S. M. Kuo, “Listening in a noisy environment: Integration of active noise control in audio products,” *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 34–43, 2016.
- [6] R. E. Bouserhal, A. Bernier, and J. Voix, “An In-Ear Speech Database in Varying Conditions of the Audio-Phonation Loop,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 1069–1077, 2019.
- [7] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Modeling of speech-dependent own voice transfer characteristics for hearables with an in-ear microphone,” *Acta Acustica*, vol. 8, 2024.
- [8] B. Gårdbæk and P. Kidmose, “On the origin of cardiovascular sounds recorded from the ear,” *IEEE Trans. Biomedical Engineering*, vol. 72, no. 1, pp. 210–216, 2025.
- [9] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, “SEANet: A multi-modal speech enhancement network,” in *Proc. Interspeech*, Shanghai, China, 2020, pp. 1126–1130.
- [10] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, “Time-Domain Multi-Modal Bone/Air Conducted Speech Enhancement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.

- [11] H. Wang, X. Zhang, and D. Wang, “Fusing Bone-Conduction and Air-Conduction Sensors for Complex-Domain Speech Enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 3134–3143, 2022.
- [12] M. Wang, J. Chen, X. Zhang, Z. Huang, and S. Rahardja, “Multi-Modal Speech Enhancement with Bone-Conducted Speech in Time Domain,” *Applied Acoustics*, vol. 200, no. 109058, 2022. Accessed: Feb. 28, 2023.
- [13] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, “Speaker Adaptation For Enhancement Of Bone-Conducted Speech,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 456–10 460.
- [14] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Multi-Microphone Noise Data Augmentation for DNN-based Own Voice Reconstruction for Hearables in Noisy Environments,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 416–420.
- [15] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, “Improving speaker discrimination of target speech extraction with time-domain speakerbeam,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 691–695.
- [16] C. Xu, W. Rao, E. S. Chng, and H. Li, “SpEx: Multi-scale time domain speaker extraction network,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.
- [17] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, “Personalized speech enhancement: New models and comprehensive evaluation,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 356–360.
- [18] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, “Neural target speech extraction: An overview,” *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.
- [19] R. Sinha, C. Rollwage, and S. Doclo, “Variants of LSTM cells for single-channel speaker-conditioned target speaker extraction,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 63, 2024.
- [20] K. Tesch and T. Gerkmann, “Insights Into Deep Non-Linear Filters for Improved Multi-Channel Speech Enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023.
- [21] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Low-complexity own voice reconstruction for hearables with an in-ear microphone,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.
- [22] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, “Asteroid: The PyTorch-based audio source separation toolkit for researchers,” in *Proc. Interspeech*, 2020, pp. 2637–2641.

- [23] J. Hauret, M. Olivier, T. Joubaud, C. Langrenne, S. Poirée, V. Zimpfer, and É. Bavu, “Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors,” *Speech Communication*, vol. 172, p. 103 238, 2025.
- [24] F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J.-H. Bach, J. Wellmann, and B. Kollmeier, “A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research,” in *Proc. AES International Conference on Headphone Technology*, 2019.
- [25] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh, and R. Aichner, “ICASSP 2023 Deep Noise Suppression Challenge,” *IEEE Open Journal of Signal Processing*, vol. 5, pp. 725–737, 2024. Accessed: Mar. 27, 2024.
- [26] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, “STFT-domain neural speech enhancement with very low algorithmic latency,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 397–410, 2023.
- [27] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. International Conference on Learning Representations*, 2015.
- [28] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – half-baked or well done?” In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.
- [29] 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,” *International Telecommunications Union*, 2001.
- [30] J. Jensen and C. H. Taal, “An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.
- [31] J. Li, K. Zhang, S. Wang, H. Li, M.-W. Mak, and K. A. Lee, “On the effectiveness of enrollment speech augmentation for target speaker extraction,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 325–332.

DISCUSSION

This chapter summarizes, discusses, and suggests topics for future research for the main contributions of Chapters 2-6 in Section 7.1, and outlines directions for further research in Section 7.2.

7.1 Chapter-by-chapter discussion

In this thesis, we developed and evaluated DNN-based OVR systems for hearables with an in-ear microphone. In noisy environments, an in-ear microphone offers the advantage of attenuating environmental noise due to occlusion of the ear canal by the hearable device. However, compared to own voice recorded by an outer microphone, in-ear own voice signals are affected by distortions such as band-limitation and low-frequency amplification. We defined these distortions as own voice transfer characteristics between the outer and the in-ear microphone.

7.1.1 *Modeling of speech-dependent own voice transfer characteristics*

In **Chapter 2**, we proposed a phoneme-dependent model of own voice transfer characteristics. The model represents the own voice transfer characteristics as a set of linear time-invariant RTFs, one for each phoneme. We considered both individual modeling, where a separate set of phoneme-specific RTFs is estimated for each talker, and talker-averaged models, where phoneme-specific RTFs are averaged across talkers. The model simulates in-ear signals while considering speech-dependent changes. Experimental results have shown that the proposed model predicts in-ear own voice signals up to 50% more accurately than speech-independent models. When predicting in-ear own voice signals for talkers that are not used for model estimation, individual models have higher prediction error than the talker-averaged model, since averaging ignores individual differences. However, the proposed modeling approach has several limitations: (i) lack of modeling individual differences of unseen talkers, (ii) limited context in temporal modeling, (iii) not considering body-produced noise, (iv) potential bias in RTF estimation, and (v) the restriction to RTF-based modeling.

First, the proposed approach is unable to model individual differences for unseen talkers. Since these differences affect in-ear own voice signals, this likely leads to

prediction errors. Related work has also considered modeling individual differences: In [1], individual DNN-based simulation of body-conducted own voice signals was performed using speaker embedding vectors. Although obtaining such vectors from unseen talkers would require recorded speech, these embeddings could be used to select own voice transfer characteristic models from a similar talker to address talker mismatch in future work. This would require speaker embeddings to contain information related to the individual differences between transfer characteristics. Future research could investigate individualization mechanisms based on speaker embeddings, and other information such as vocal tract length, sex, or pitch of the talker could improve modeling accuracy. It may also be useful to group talkers based on physiological parameters to obtain averaged models of similar talkers. A similar approach based on grouping similar talkers was proposed for personalized speech codecs in [2].

Second, temporal modeling is limited, since the current frame-based approach assumes stationarity within each frame, neglecting phoneme transitions in the process. Although the temporal smoothing in (2.17) reduces discontinuities, it may not accurately reflect real phoneme transitions. It may be beneficial to include context in the form of phonetic units like triphones or larger semantic units like words. The current approach also does not account for pronunciation differences of the same word in different sentence contexts (e.g., in a statement compared to a question), as it assumes identical transfer characteristics for all occurrences of each phoneme.

Third, the proposed model does not account for body-produced noise, which is uncorrelated with own voice and therefore difficult to predict. The absence of body-produced noise in the simulated in-ear own voice signals leads to errors when trying to predict recorded in-ear own voice signals, which always contain body-produced noise. In [3], randomly selected body-produced noise signals were added to simulated in-ear own voice signals. Similarly, the noise data augmentation technique proposed in Appendix A adds white noise with a low amplitude to the simulated in-ear environmental noise signal to coarsely approximate body-produced noise. Although this technique is used in Chapters 3-5, the influence of this approximation was not investigated. White noise is likely not a realistic approximation for body-produced noise (which consists of mostly low-frequency content). To achieve a more realistic simulation, future work should investigate matching the spatial coherence patterns between recorded outer and in-ear own voice signals by adding synthetic or augmented body-produced noise to simulated in-ear own voice signals. Since body-produced noise relates to physiological processes [4], parameters like heart rate of the device user could be included in the simulation.

Fourth, the least squares approach for RTF estimation in (3.7) may not be optimal, since that own voice is a suboptimal excitation signal with limited bandwidth and poor repeatability. In addition, body-produced noise and microphone self-noise may bias the transfer function estimate. Methods that use measurement noise power spectral densities for bias correction or combine estimation methods in a frequency-dependent manner [5] could improve phoneme-dependent RTF estimation and potentially lead to higher simulation accuracy.

Finally, different approaches for own voice data augmentation based on DNNs [1, 6–8] or a convolutional transfer function model [9] have been proposed to train OVR systems. Future research should investigate these approaches and compare them to the proposed speech-dependent approach in terms of modeling accuracy.

7.1.2 *Speech-dependent data augmentation for OVR*

In **Chapter 3**, we proposed to perform data augmentation using the phoneme-dependent own voice transfer characteristic models from Chapter 2. Specifically, we used a system based on the FT-JNF architecture [10] which computes complex-valued masks for the outer and the in-ear microphone signals. In the experimental evaluation, this system outperformed the EBEN baseline system (trained in a fully supervised manner), which only used the in-ear microphone. Results show that training with phoneme-dependent individual augmentation followed by additional fine-tuning with recorded signals yields the best performance, achieving improvements of about 1.3 in terms of PESQ and over 0.15 in terms of STOI compared to the noisy outer microphone signal. Comparing phoneme-dependent individual models with phoneme-dependent talker-averaged models for augmentation, we observed better performance with individual models. While the results demonstrated that phoneme-dependent individual augmentation can substantially reduce training data requirements for the considered FT-JNF-based OVR system, the evaluation is restricted in several respects: Different data augmentation techniques introducing different amounts of variance into the training data, assessing generalization only across recorded talkers using a single device and auxiliary sensor type, and only for one specific DNN architecture. In the following, we discuss these generalization aspects: (i) individual differences, (ii) random variance, (iii) the chosen network architecture, (iv) generalization across devices and auxiliary sensors, and (v) better simulation and fine-tuning to extend the proposed augmentation scheme. After discussing these aspects, we suggest other possible directions for future research.

First, while talker-averaged models led to higher modeling accuracy for predicting in-ear own voice signals of unseen talkers (Chapter 2), individual models led to better OVR performance. It is likely that the additional variance in the training data from using multiple individual models instead of a single talker-averaged model improves OVR performance. Even in-ear own voice signals simulated using one talker-averaged model are less accurate than in-ear own voice signals simulated using multiple individual models, they appear to improve the capabilities of an OVR system trained with simulated signals to generalize to recorded signals.

Second, an increase in performance due to additional variance was also observed when comparing speech-dependent data augmentation to random-phoneme data augmentation, suggesting that while about two-thirds of the PESQ improvement can be attributed to introducing random variance to the training data, the remaining improvement is due to modeling of speech-dependent behavior. Further increasing variance in addition to speech-dependent individual augmentation may improve generalization capabilities to recorded own voice signals. Further research could incorporate random changes in the RTF magnitudes, as in [11], or random se-

lection of RTFs estimated from different utterances, as in [3]. Randomly varying key parameters of the transfer characteristics, such as cutoff frequency or low-frequency gain, which depend on the device or insertion quality, might improve performance on recorded signals further or enable generalization between devices. Instead of incorporating these variations, removing them from the DNN input signals (e.g., using a low-pass filter for the body-conducted signal as in [7], or by device-specific equalization filtering) could also improve generalization to recorded signals.

Third, Chapter 3 did not consider that different DNN architectures may be better suited for generalization to recorded own voice signals or to different devices. For example, the filter-and-sum mechanism used in the FT-JNF architecture (see Figure 3.3) may be less suitable for devices with changing array geometries, such as with a close-talk microphone on an adjustable arm in [12]. It would be interesting for future research to compare the FT-JNF architecture with other architectures specifically proposed for OVR, such as [13, 14].

Fourth, it would be interesting to investigate data augmentation for OVR across different devices and sensors. For instance, the large Vibravox dataset of own voice recordings [12] containing parallel recordings of own voice with multiple auxiliary sensors could be used for a systematic benchmark. These recordings could also enable a fair comparison between training with a large dataset of augmented signals and training with an equally sized dataset of recorded signals to investigate generalization to (unseen) recorded signals. Although imperfect simulation of in-ear own voice signals may result in lower performance than training with a sufficient amount of recorded signals, the increased variance from speech-dependent augmentation might compensate for this. It may also be sufficient to use large amounts of augmented data even if the simulation is not perfectly realistic. Similarly, future research could investigate whether the proposed simulation methods and OVR approaches generalize to throat microphones, for example using the recent TAPS dataset [15].

Fifth, it is possible that better simulation of in-ear own voice signals or different fine-tuning strategies would improve OVR generalization to recorded own voice signals. In future research it would be interesting to compare DNN-based techniques [6–8, 16], or the convolutional transfer function-based model from [9] to simulate in-ear own voice signals with the proposed speech-dependent data augmentation technique. Comparing these approaches in terms of simulation accuracy (as suggested in Section 7.1.1) may not reflect the efficacy of the approaches for training OVR systems, as observed for the talker-averaged approaches. DNN-based techniques may also be able to simulate time-varying changes or individual own voice transfer characteristics better than the data augmentation proposed in this thesis, potentially resulting in more accurate simulation or greater variance in the augmented training data. Also, substantial performance gains are often achieved by additional fine-tuning after training with simulated signals. The gain from fine-tuning compared to only training with simulated signals could also reflect limited generalization ability, since fully supervised discriminative speech enhancement approaches are known to sometimes overfit to dataset characteristics. Overfitting was observed for e.g., the frequency response of the microphone with which the training

dataset was recorded [17], or in case of bandwidth extension, the filter the training data was band-limited with [18]. In analogy, an OVR system trained on simulated data may overfit to the own voice transfer characteristic models used for augmentation. In this case, it would be especially interesting for further research to improve modeling accuracy.

Finally, rather than improving in-ear own voice modeling accuracy to recreate a 'digital twin' of the considered device, future research could focus on speech enhancement approaches with better generalization across different subproblems of speech enhancement. For example, generative approaches to universal speech enhancement have shown promising generalization results (see e.g., [19–22]). Future research on general speech restoration approaches should compare their performance against approaches designed only OVR.

7.1.3 *Low-complexity OVR for hearables*

In **Chapter 4**, we investigated low-complexity variants of the OVR system from Chapter 3, trained using speech-dependent data augmentation and fine-tuning with recorded signals. Compared to the complexity of the OVR system considered in the previous chapter (1.39 M parameters, 22.38 GMACs/s), we have heavily reduced computational complexity (down to 13 k parameters, 0.23 GMACs/s). The results showed that even with low computational complexity and few recorded own voice signals, the low-complexity variants considerably improve speech quality according to objective metrics. The proposed low-complexity variants performed better than baseline systems with higher or comparable complexity [23–26].

While Chapter 4 primarily targeted computational complexity and data requirements, several additional practical aspects remain unaddressed. Future research could investigate quantization and pruning of DNN parameters [27–29] to further reduce complexity. Moreover, considering applications in which the reconstructed own voice is transmitted to other devices, it may be beneficial to include transmission properties such as codec and post-transmission bandwidth during training. Including these distortions could relax system requirements and improve post-transmission quality, which is more relevant to the listener than pre-transmission quality.

Beyond compressing the model and matching the transmission channel, architectural choices from low-complexity speech enhancement may also improve low-complexity OVR architectures. The limited bandwidth of own voice recorded at the in-ear microphone could be exploited by focussing on low-frequency processing to reduce computational complexity, as in [30]. In addition, some low-complexity speech enhancement systems reduce input dimensionality via filterbanks (e.g., equivalent rectangular bandwidth (ERB) filterbanks [31, 32]) and perform efficient sub-band processing. This approach could be adapted to the FT-JNF architecture, although modifications to the F-LSTM may be required to facilitate non-uniform step sizes across sub-bands with different bandwidths.

Given that most processing operations in the FT-JNF architecture are performed by the LSTM layers, another direction could be to replace LSTMs with more efficient recurrent layers, such as gated recurrent units (GRUs) [33, 34]. In [34], it was also observed for a GRU-based single-channel speech enhancement system with low complexity that a real-valued mask performed similarly to a complex-valued mask. Future research could examine whether this also applies to low-complexity OVR.

Finally, model compression and architectural simplifications can sometimes reduce performance in ways that are hard to compensate for through standard training. In this context, knowledge distillation from a larger, better-performing DNN could be used to guide the training of a smaller, low-complexity DNN. In this approach, the output of intermediate layers of a larger trained DNN are used to train the low-complexity DNN by adding additional loss terms into the training, as shown for speech enhancement in [35, 36]. This approach could also improve the performance of low-complexity OVR systems without increasing their complexity, potentially enabling better quality-complexity trade-offs compared to standard training.

7.1.4 *Subjective quality evaluation of personalized OVR systems*

In **Chapter 5**, we proposed to apply the speech-dependent data augmentation technique from Chapter 3 to training-based personalization of OVR systems to individual talkers, using either personalized data augmentation, personalized fine-tuning, or a combination of both. We conducted a formal subjective MUSHRA-like listening test with $N = 25$ normal-hearing subjects to evaluate the benefit of generic (non-personalized) and personalized OVR systems, and compared objective metric predictions with subjective quality ratings of unprocessed and reconstructed own voice signals. The listening test results show that our DNN-based OVR systems (using both the outer and the in-ear microphone) substantially improved subjective quality compared to the unprocessed (noisy) outer and in-ear microphone signals and several baseline OVR systems. Although objective metrics predicted performance gains for personalized compared to generic (non-personalized) processing, these improvements were only observed for some conditions in the subjective ratings. Between different talkers, large differences in predicted quality were observed, while subjective ratings of two exemplary talkers were rather similar to each other. Contrary to the predictions of objective metrics, subjective ratings indicate that personalized processing only yields additional benefits in some cases. Importantly, most objective metrics underestimated the quality of noisy in-ear own voice signals and overestimated the quality of bandwidth-extended in-ear own voice signals compared with subjective ratings. However, several metrics showed a high correlation with subjective ratings, especially the intrusive ESTOI and GPSM^a and the non-intrusive LEAP.

Several factors may explain why the proposed training-based personalization did not consistently improve quality compared to generic OVR systems. First, the improvements indicated by objective metrics might be too small to be perceptible by human listeners. Second, personalized data augmentation only considers the own voice transfer characteristics of the target talker, while the input speech signals

used for simulating in-ear own voice signals originate from many different talkers (not even including the target talker). As a result, individual speech characteristics such as pitch and timbre of the target talker are not reflected in the augmented training data. If enough own voice recordings from the target talker were available, the benefit of training-based personalization could be greater. Experimental results in Chapter 5 support this: In terms of both objective metrics and subjective ratings, OVR systems using personalized fine-tuning with recorded own voice signals of the target talker performed better than systems using only personalized data augmentation. A promising extension could be to combine personalized data augmentation with personalized speech synthesis [37]. Third, the acoustic scenarios considered in Chapter 5 may not be ideal to expose the full potential of personalization. For own voice in environmental noise without interfering talkers, generic OVR systems using both an outer and an in-ear microphone may suffice since the cues provided by the in-ear microphone (e.g., information on frequency-dependent own voice presence up to the cutoff frequency) may be sufficient to distinguish own voice from environmental noise, leaving limited headroom for additional personalization benefits. In contrast, previous research in [6] and our work in Chapter 6 showed that the benefits of an in-ear microphone for personalized processing are especially large when interfering talkers are present.

The findings in Chapter 5 also highlight why subjective evaluation remains important for OVR. The listening test showed that not every improvement predicted by objective metrics is reflected by subjective ratings, indicating that subjective evaluations are better suited for comparing OVR systems. However, this is not always practical, e.g., when comparing many trained systems or when only small performance differences are expected. In such cases, objective metrics remain valuable, e.g., to tune hyperparameters, compare different DNN architectures or compare different pre-processing pipelines for the training data. Based on the correlation analysis in Chapter 5, we would hence recommend ESTOI, GDSM^a, and LEAP for such optimization. Still, the estimation biases observed for noisy band-limited and processed bandwidth-extended in-ear own voice signals in the subjective evaluation should be kept in mind when interpreting these metrics. This points to a broader limitation: None of the metrics were specifically designed to predict OVR performance, but rather for broadband speech enhancement or bandwidth extension without noise reduction (e.g., WV-MOS). Although noisy in-ear own voice signals are clearly distorted due to band-limitation and low-frequency attenuation, listeners do not rate these effects as negatively as additive noise in the outer microphone signal. Moreover, comparing noisy outer and in-ear microphone signals at the same environmental noise level is inherently biased by their different SNR, which favors the in-ear microphone signal. However, this SNR advantage reflects real-world conditions and is therefore relevant. To better disentangle these factors, future research could use multi-dimensional audio quality ratings, e.g., [38]. Similarly, it could be interesting to include an anchor signal that adds noise to the in-ear microphone signal to match the SNR of the outer microphone signal, thereby isolating the perceptual impact of SNR benefits versus own voice distortions.

Further, future research could compare training-based personalization using own voice transfer characteristics with other personalization methods (discussed in Section 7.1.5 in more detail).

7.1.5 PAS-SE: *Personalized auxiliary-sensor speech enhancement*

Whereas Chapter 5 proposed training-based personalization, in **Chapter 6**, we explored personalization via enrollment utterances for OVR systems using an outer and an in-ear microphone. Enrollment-based personalization only requires a single enrollment utterance of the talker recorded with the in-ear microphone. For enrollment-based personalization, we add a speaker encoder branch to the FT-JNF architecture considered in previous chapters. The architecture used in this chapter performs single-channel real-valued masking of the outer microphone signal instead of the filter-and-sum mechanism using multiple complex-valued masks considered in previous chapters. We also proposed several data augmentation techniques (referred to as training-time augmentation configurations in Chapter 6) that enable the training of multi-channel OVR systems (referred to as AS-SE systems in Chapter 6) without the need for recorded in-ear interferer signals. We conducted an experimental evaluation using the Vibravox dataset [12] considering environmental noise and interfering talkers. Results demonstrate that enrollment-based personalization using an in-ear microphone is very effective in such scenarios, achieving up to 10 dB SI-SDR improvement over unprocessed signals and outperforming generic (non-personalized) OVR systems, especially in scenarios with interfering talkers. Moreover, the experimental results demonstrate that the proposed augmentation techniques enable interfering talker suppression and improve cross-dataset generalization, as shown by additional experiments using the Hearpiece dataset considered in previous chapters (referred to as the Oldenburg dataset in Chapter 6) for testing. Personalized OVR systems were shown to remain robust when using noisy in-ear enrollment utterances.

In terms of objective metrics, the benefits of enrollment-based personalization are substantially larger than the benefits of training-based personalization observed in Chapter 5. This motivates a systematic comparison of personalization strategies not only across noise and interferer conditions, but also with respect to the amount and type of recorded data required for personalization. However, enrollment-based personalization introduces practical challenges that need to be weighed against its benefits: First, collecting enrollment utterances in real applications requires additional control structures, such as explicit user interaction or automatic enrollment selection. Second, enrollment-based personalization increases DNN size due to the speaker encoder branch. If encoding is performed on the hearable device, additional resources are required for computation and storage. If encoding is performed on another device (e.g., a smartphone or remote server), additional infrastructure is required to support communication between the hearable and the external device. These practical challenges are key considerations when evaluating the feasibility of enrollment-based personalized OVR in real products.

Beyond the personalization strategy itself, Chapter 6 also provided insights into architectural choices that affect cross-dataset generalization. In particular, the system using magnitude STFT inputs and a single real-valued mask for the outer microphone generalized well from the Vibravox dataset to the Hearpiece dataset. One possible explanation is that this architecture relies less on device-specific properties of the in-ear signals than the filter-and-sum architecture used in other chapters. For example, phase differences between the in-ear microphone and the close-talk reference microphone can vary in the Vibravox dataset because the close-talk microphone is positioned on an adjustable arm, unlike the fixed-position outer microphone in the Hearpiece dataset. More generally, other factors such as effective sensor bandwidth may also play a role in cross-dataset or cross-sensor generalization. This interpretation is supported by the Speakerbeam baseline systems, which did not generalize well to a different dataset, likely because their learned filterbanks overfit to device-specific factors. Since hearable devices differ widely in shape, inter-microphone distances, and resulting relative transfer functions, this raises an important question: Should own voice pickup be addressed with a device-specific OVR system or an OVR system that generalizes across device form factors? Understanding the cross-device generalization of different DNN architectures for OVR therefore remains an important topic for future research.

In addition to generalization across datasets and devices, robustness to practical enrollment conditions is another important issue. In Chapter 6, personalization with in-ear enrollment utterances was shown to be more robust to environmental noise in the enrollment utterance than personalization with outer microphone enrollment utterances. Training with noisy enrollment signals may further increase the robustness for both in-ear and outer microphone enrollment. More broadly, future work could investigate robustness to other constraints, such as enrollment length, or enrollment signals collected with different devices (e.g., other hearables or smartphones). Related to Chapter 4, future work could also investigate whether enrollment-based personalization is still effective for low-complexity variants.

The results in Chapter 6 showed that while the augmentation techniques enabling interferer suppression for multi-channel OVR systems do not accurately model interferer transmission to the in-ear microphone, they consistently improve performance in terms of interferer suppression. This contrasts with the findings in Appendix A, where more accurate modeling of noise transmission from the outer to the in-ear microphone led to better OVR performance in terms of noise reduction. Taken together, these results suggest that more accurate modeling of in-ear interferer transmission could further improve interferer suppression. In addition, personalized systems (PAS-SE) appear to benefit less from approximating the in-ear interferer components using a broadband gain during training than non-personalized systems (AS-SE). This could suggest that personalization increases robustness to in-ear interferer variability. Further research should therefore systematically compare augmentation configurations for in-ear noise and interferer components for both generic (non-personalized) and personalized approaches.

7.2 General discussion

While the experiments reported in this thesis focused on DNN architectures based on the FT-JNF architecture [10], many alternative architectures are potentially interesting for OVR. Prior work has explored various DNN architectures for this task, including fully convolutional networks [39], convolutional UNet-like structures [3, 14, 30], recurrent neural networks [40, 41], transformers [42], attention mechanisms [43, 44], and selective state space models [45]. Although many of these OVR architectures were inspired by recent advances in speech enhancement, it remains unclear whether OVR requires specialized architectures, or if general speech enhancement architectures are sufficient. A key obstacle to answering this question is that existing studies are often difficult to compare directly. As Tables 1.1 and 1.2 show, approaches differ not only in the number of parameters, but also in training and test data and training paradigms. Therefore, future research should benchmark different architectures under realistic conditions to identify which design choices and training paradigms are most beneficial for OVR, especially considering computational complexity and processing latency. Similarly, it would be interesting to compare DNN-based own voice pickup with traditional methods for multi-microphone enhancement, such as beamforming with compact earbud microphone arrays, or to investigate hybrid approaches that combine DNN-based voice pickup with beamforming. Such a benchmark should also cover practical edge cases that may strongly affect in-ear or outer microphone signals. Examples include physical exercise [46], which increases body-produced noise at the in-ear microphone, or wind noise [9, 47, 48], which increases environmental noise and can even lead to destructive interference at the outer microphone. Beyond acoustic conditions, device factors should be represented as well. Although the placement of microphones or auxiliary sensors is often constrained by industrial design, a benchmark could still include multiple devices with different sensor locations, e.g., comparing outer device microphones with close-talk microphones or including other locations for contact microphones, to better understand how sensor placement influences OVR performance and robustness. Related to sensor choice, another open question is whether OVR approaches developed for body-conduction or contact microphones, accelerometers, or similar auxiliary sensors remain effective when applied to in-ear microphones, which also pick up environmental noise.

In practice, OVR systems in hearables are also unlikely to operate in isolation. In-ear microphones are often used for other device features, such as active noise cancellation or active occlusion cancellation, which can alter the recorded in-ear microphone signal due to active playback of a cancellation signal. Hearables are also often used to play music or the voice of a remote communication partner, in which case the loudspeaker signals need to be considered in a joint speech enhancement and feedback cancellation scheme (see e.g., [49, 50]). As a result, in practice OVR likely needs to be integrated with other device features to prevent artifacts, noise leakage, and feedback. Such integration however also creates opportunities, as information from these features, such as voice activity or transfer path estimates, could be incorporated into the OVR system. For example, indicators of device fit

quality could help determine the utility of the in-ear microphone signal, which could be exploited in signal-dependent sensor fusion strategies.

Finally, since voice pickup benefits remote listeners rather than the device user, its advantages may be hard to communicate to potential users. For remote listeners, it may not be obvious whether improved speech quality is due to their own device or the sender's device. Because this feature is hence difficult to demonstrate and communicate, the trade-off between device resources and improved user experience needs to be carefully weighed.

Even when this chapter has discussed many open questions and practical challenges, we hope that this thesis will inspire and support future research and development on own voice reconstruction.

References

- [1] M. Pucher and T. Woltron, "Conversion of Airborne to Bone-Conducted Speech with Deep Neural Networks," in *Proc. Interspeech*, Brno, Czechia, Aug. 2021. DOI: [10.21437/Interspeech.2021-473](https://doi.org/10.21437/Interspeech.2021-473).
- [2] I. Jang, H. Yang, W. Lim, S. Beack, and M. Kim, "Personalized neural speech codec," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 991–995. DOI: [10.1109/ICASSP48485.2024.10446067](https://doi.org/10.1109/ICASSP48485.2024.10446067).
- [3] M. Ohlenbusch, C. Rollwage, and S. Doclo, "Training strategies for own voice reconstruction in hearing protection devices using an in-ear microphone," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022. DOI: [10.1109/IWAENC53105.2022.9914801](https://doi.org/10.1109/IWAENC53105.2022.9914801).
- [4] B. Gårdbæk and P. Kidmose, "On the origin of cardiovascular sounds recorded from the ear," *IEEE Trans. on Biomedical Engineering*, vol. 72, no. 1, pp. 210–216, 2025. DOI: [10.1109/TBME.2024.3445412](https://doi.org/10.1109/TBME.2024.3445412).
- [5] M. Blau, R. Roden, N. Hauenschild, S. Kersten, R. Rehman, M. Vorländer, and J. Fels, "Methods to experimentally characterize the own-voice-generated objective occlusion effect induced by hearables," *Acta Acustica*, vol. 9, p. 73, 2025. DOI: [10.1051/aacus/2025055](https://doi.org/10.1051/aacus/2025055).
- [6] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "SEANet: A multi-modal speech enhancement network," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1126–1130. DOI: [10.21437/Interspeech.2020-1563](https://doi.org/10.21437/Interspeech.2020-1563).
- [7] H. Wang, X. Zhang, and D. Wang, "Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 3134–3143, 2022. DOI: [10.1109/TASLP.2022.3209943](https://doi.org/10.1109/TASLP.2022.3209943).
- [8] D. Ma, T. Dang, M. Ding, and R. Balan, "Clearspeech: Improving voice quality of earbuds using both in-ear and out-ear microphones," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 4, Jan. 2024. DOI: [10.1145/3631409](https://doi.org/10.1145/3631409).

- [9] J. Heitkaemper, J. Caroselli, M. McKinnon, A. Narayanan, and N. Howard, “Bone conducted signal guided speech enhancement for voice assistant on earbuds,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025. DOI: [10.1109/ICASSP49660.2025.10889416](#).
- [10] K. Tesch and T. Gerkmann, “Insights into deep non-linear filters for improved multi-channel speech enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023. DOI: [10.1109/TASLP.2022.3221046](#).
- [11] L. He, H. Hou, S. Shi, X. Shuai, and Z. Yan, “Towards bone-conducted vibration speech enhancement on head-mounted wearables,” in *Proc. Annual International Conference on Mobile Systems, Applications and Services*, New York, USA, Jun. 2023, pp. 14–27. DOI: [10.1145/3581791.3596832](#).
- [12] J. Hauret, M. Olivier, T. Joubaud, C. Langrenne, S. Poirée, V. Zimpfer, and É. Bavu, “Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors,” *Speech Communication*, Apr. 2025. DOI: [10.1016/j.specom.2025.103238](#).
- [13] K. Kuang, F. Yang, and J. Yang, “A lightweight speech enhancement network fusing bone- and air-conducted speech,” *J. Acoust. Soc. Am.*, vol. 156, no. 2, pp. 1355–1366, Aug. 2024. DOI: [10.1121/10.0028339](#).
- [14] C. Li, F. Yang, and J. Yang, “Bone conduction-aided speech enhancement with two-tower network and contrastive learning,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 33, pp. 163–174, 2025. DOI: [10.1109/TASLP.2024.3512207](#).
- [15] Y. Kim, Y. Song, and Y. Chung, *TAPS: Throat and acoustic paired speech dataset for deep learning-based speech enhancement*, Feb. 2025. DOI: [10.48550/arXiv.2502.11478](#). [Online]. Available: <https://arxiv.org/abs/2502.11478>.
- [16] M. Pucher and T. Woltron, “Conversion of Airborne to Bone-Conducted Speech with Deep Neural Networks,” in *Proc. Interspeech*, Brno, Czechia, Aug. 2021, pp. 1–5. DOI: [10.21437/Interspeech.2021-473](#).
- [17] A. Pandey and D. Wang, “On cross-corpus generalization of deep learning based speech enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 2489–2499, 2020. DOI: [10.1109/TASLP.2020.3016487](#).
- [18] H. Wang and D. Wang, “Towards robust speech super-resolution,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 2058–2066, 2021. DOI: [10.1109/TASLP.2021.3054302](#).
- [19] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, “Universal speech enhancement with score-based diffusion,” *arXiv:2206.03065*, Jun. 2022. DOI: [10.48550/arXiv.2206.03065](#).

- [20] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Causal diffusion models for generalized speech enhancement," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 780–789, 2024. DOI: 10.1109/OJSP.2024.3379070.
- [21] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann, "Diffusion models for audio restoration: A review," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 72–84, 2024. DOI: 10.1109/MSP.2024.3445871.
- [22] W. Zhang, K. Saijo, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Kumar, M. Sach, W. Wang, Y. Fu, S. Watanabe, T. Fingscheidt, and Y. Qian, "Lessons Learned from the URGENT 2024 Speech Enhancement Challenge," in *Proc. Interspeech*, Rotterdam, The Netherlands, Aug. 2025, pp. 853–857. DOI: 10.21437/Interspeech.2025-1246.
- [23] M. Ohlenbusch, C. Rollwage, and S. Doclo, "Training strategies for own voice reconstruction in hearing protection devices using an in-ear microphone," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022. DOI: 10.1109/IWAENC53105.2022.9914801.
- [24] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, "EBEN: Extreme bandwidth extension network applied to speech signals captured with noise-resilient body-conduction microphones," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, Jun. 2023. DOI: 10.1109/ICASSP49357.2023.10096301.
- [25] M. Ohlenbusch, C. Rollwage, and S. Doclo, "Speech-dependent data augmentation for own voice reconstruction with hearable microphones in noisy environments," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 8, no. 32, 2025. DOI: 10.1186/s13636-025-00418-1.
- [26] 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, Dec. 2023. DOI: 10.1109/OJSP.2023.3343320.
- [27] J.-Y. Wu, C. Yu, S.-W. Fu, C.-T. Liu, S.-Y. Chien, and Y. Tsao, "Increasing compactness of deep learning based speech enhancement models with parameter pruning and quantization techniques," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1887–1891, 2019. DOI: 10.1109/LSP.2019.2951950.
- [28] M. Stamenovic, N. L. Westhausen, L.-C. Yang, C. Jensen, and A. Pawlicki, "Weight, Block or Unit? Exploring Sparsity Tradeoffs for Speech Enhancement on Tiny Neural Accelerators," in *Proc. NeurIPS Workshop Efficient Natural Language and Speech Processing*, Nov. 2021. DOI: 10.48550/arXiv.2111.02351.
- [29] K. Tan and D. Wang, "Towards model compression for deep learning based speech enhancement," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 1785–1794, 2021. DOI: 10.1109/TASLP.2021.3082282.

- [30] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, “Configurable EBEN: Extreme bandwidth extension network to enhance body-conducted speech capture,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 3499–3512, 2023. DOI: 10.1109/TASLP.2023.3313433.
- [31] H. Schröter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, “DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7407–7411. DOI: 10.1109/ICASSP43922.2022.9747055.
- [32] X. Rong, T. Sun, X. Zhang, Y. Hu, C. Zhu, and J. Lu, “GTCRN: A speech enhancement model requiring ultralow computational resources,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 971–975. DOI: 10.1109/ICASSP48485.2024.10448310.
- [33] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, “Towards Efficient Models for Real-Time Deep Noise Suppression,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 656–660. DOI: 10.1109/ICASSP39728.2021.9413580.
- [34] R. Sinha, C. Rollwage, and S. Doclo, “Low-complexity real-time single-channel speech enhancement based on skip-GRUs,” in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, Sep. 2023, pp. 181–185. DOI: 10.30420/456164035.
- [35] R. D. Nathoo, M. Kegler, and M. Stamenovic, “Two-step knowledge distillation for tiny speech enhancement,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 141–10 145. DOI: 10.1109/ICASSP48485.2024.10446796.
- [36] R. Metzger, M. Ohlenbusch, C. Rollwage, and S. Doclo, “Comparison of knowledge distillation methods for low-complexity multi-microphone speech enhancement using the FT-JNF architecture,” in *Proc. ITG Conference on Speech Communication*, Berlin, Germany, 2025, pp. 131–135.
- [37] A. Kuznetsova, A. Sivaraman, and M. Kim, “The potential of neural speech synthesis-based data augmentation for personalized speech enhancement,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 2023. DOI: 10.1109/ICASSP49357.2023.10096601.
- [38] B. Naderi, R. Cutler, and N.-C. Ristea, “Multi-dimensional speech quality assessment in crowdsourcing,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 696–700. DOI: 10.1109/ICASSP48485.2024.10447225.
- [39] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, “Time-domain multi-modal bone/air conducted speech enhancement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020. DOI: 10.1109/LSP.2020.3000968.

- [40] H. Q. Nguyen and M. Unoki, "Improvement in bone-conducted speech restoration using linear prediction and long short-term memory model," *Journal of Signal Processing*, vol. 24, no. 4, pp. 175–178, 2020. DOI: [10.2299/jsp.24.175](#).
- [41] M. Ohlenbusch, C. Rollwage, and S. Doclo, "Low-complexity own voice reconstruction for hearables with an in-ear microphone," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025. DOI: [10.1109/ICASSP49660.2025.10887874](#).
- [42] C. Zheng, L. Xu, X. Fan, J. Yang, J. Fan, and X. Huang, "Dual-path transformer-based network with equalization-generation components prediction for flexible vibrational sensor speech enhancement in the time domain," *J. Acoust. Soc. Am.*, vol. 151, no. 5, pp. 2814–2825, Apr. 2022. DOI: [10.1121/10.0010316](#).
- [43] H. Wang, X. Zhang, and D. Wang, "Attention-based fusion for bone-conducted and air-conducted speech enhancement in the complex domain," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7757–7761. DOI: [10.1109/ICASSP43922.2022.9746374](#).
- [44] C. Li, F. Yang, and J. Yang, "Restoration of bone-conducted speech with U-Net-like model and energy distance loss," *IEEE Signal Process. Lett.*, vol. 31, pp. 166–170, 2024. DOI: [10.1109/LSP.2023.3347149](#).
- [45] Y. Sui, M. Zhao, J. Xia, X. Jiang, and 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 *Proc. ACM Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, New York, USA, Nov. 2024. DOI: [10.1145/3699757](#).
- [46] F. Han, P. Yang, Y. Zuo, F. Shang, F. Xu, and X.-Y. Li, "Earspeech: Exploring in-ear occlusion effect on earphones for data-efficient airborne speech enhancement," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 3, Sep. 2024. DOI: [10.1145/3678594](#).
- [47] S. Franz and J. Bitzer, "Multi-channel algorithms for wind noise reduction and signal compensation in binaural hearing aids," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.
- [48] D. Mirabilii, "Wind noise analysis, synthesis, and reduction for speech enhancement using compact microphone arrays," PhD thesis, Technische Fakultät, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2023.
- [49] N. C. Ristea, E. Indenbom, A. Saabas, T. Pärnamaa, J. Guzhvin, and R. Cutler, "DeepVQE: Real time deep voice quality enhancement for joint acoustic echo cancellation, noise suppression and dereverberation," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 3819–3823. DOI: [10.21437/Interspeech.2023-1028](#).

- [50] E. Seidel, P. Mowlae, and T. Fingscheidt, “Efficient high-performance bark-scale neural network for residual echo and noise suppression,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 1386–1390. DOI: [10.1109/ICASSP48485.2024.10446427](https://doi.org/10.1109/ICASSP48485.2024.10446427).

CONCLUSION

In this thesis, we have developed and evaluated OVR systems for hearables with an in-ear microphone. In-ear microphones offer distinct advantages for own voice pickup since the occlusion of the ear canal attenuates environmental sounds at the microphone. At the same time, in-ear own voice signals are affected by characteristic distortions, most notably bandwidth limitation and low-frequency amplification. Crucially, these distortions are not constant, as these transfer characteristics vary with speech production and differ between individuals. This variability poses major challenges for DNN-based OVR, where the availability of representative training data and generalization across different talkers are key bottlenecks.

To address these challenges, this thesis makes the following three main contributions to the field of own voice reconstruction: First, we proposed a speech-dependent model of own voice transfer characteristics that explicitly accounts for phoneme-related changes in the transmission of own voice from the outer to the in-ear microphone. By incorporating speech-dependent behavior, the proposed model improves the prediction of in-ear own voice signals compared to speech-independent approaches and provides the basis for simulating realistic in-ear recordings. Second, we demonstrated how models of own voice transfer characteristics can be used for speech-dependent own voice data augmentation, enabling the simulation of large in-ear own voice datasets from limited recorded signals. This directly addresses a central challenge in DNN-based OVR, namely limited availability of paired recordings with the outer and the in-ear microphone. In addition, we investigated low-complexity OVR system variants that substantially reduce computational requirements while maintaining moderate processing latency. Third, we conducted a formal subjective quality evaluation of generic and personalized OVR systems and analyzed the relationship between subjective ratings and objective metric predictions. While the evaluation revealed strong benefits from OVR processing, personalized processing did not consistently lead to higher subjective quality, revealing limitations of objective metrics for system comparison. By quantifying these discrepancies, we identified objective measures that are particularly suitable for OVR evaluation. Beyond training-based personalization, we further proposed enrollment-based personalization for OVR systems, achieving large improvements for the suppression of interfering talkers.

Overall, this thesis demonstrates how the unique advantages of in-ear microphones can be exploited to improve own voice reconstruction, and it addresses key practical requirements for OVR systems: modeling and simulation of in-ear own voice signals, data-efficient training through augmentation, computational feasibility, and perceptually grounded evaluation.

Nevertheless, several limitations remain and motivate future research. Modeling of own voice transfer characteristics can be improved further by more advanced estimation and simulation methods, including better handling of temporal dynamics, body-produced noise, and measurement bias. Such improvements might also be beneficial for data augmentation, even though it remains an open question how large the relative importance of realistic simulation compared to increasing the variance of the training data. In addition, the subjective evaluation of personalized OVR in this thesis focused on scenarios with environmental noise without interfering talkers, whereas other experiments showed that personalization shows the largest benefits for interferer suppression. Future research should systematically benchmark OVR approaches designed for in-ear microphones with those designed for other auxiliary sensors under controlled, realistic conditions including different devices, acoustic edge cases, and practical constraints, to clarify which architectures and training paradigms achieve the highest quality under these conditions.



MULTI-MICROPHONE NOISE DATA AUGMENTATION FOR DNN-BASED OWN VOICE RECONSTRUCTION FOR HEARABLES IN NOISY ENVIRONMENTS

This appendix is identical in content to the publication: M. Ohlenbusch, C. Rollwage, and S. Doclo, “Multi-microphone noise data augmentation for DNN-based own voice reconstruction for hearables in noisy environments,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 416–420. DOI: [10.1109/ICASSP48485.2024.10447066](https://doi.org/10.1109/ICASSP48485.2024.10447066).

Authors	Author’s contribution							
	A	B	C	D	E	F	G	H
Mattes Ohlenbusch	✗	✗	✗	✗	✗	✗	✗	✗
Christian Rollwage	✗			✗		✗	✗	✗
Simon Doclo	✗			✗		✗	✗	✗

- A - Substantial contributions to the conception or design of the work
- B - Acquisition of the data
- C - Analysis of the data
- D - Interpretation of the data
- E - Drafting the work
- F - Revising the work critically
- G - Final approval of the version to be published
- H - Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Abstract

Hearables with integrated microphones may offer communication benefits in noisy working environments, e.g. by transmitting the recorded own voice of the user. Systems aiming at reconstructing the clean and full-bandwidth own voice from noisy microphone recordings are often based on supervised learning. Recording a sufficient amount of noise required for training such a system is costly since noise transmission between outer and inner microphones varies individually. Previously proposed methods either do not consider noise, only consider noise at outer microphones or assume inner and outer microphone noise to be independent during training, and it is not yet clear whether individualized noise can benefit the training of and own voice reconstruction system. In this paper, we investigate several noise data augmentation techniques based on measured transfer functions to simulate multi-microphone noise. Using augmented noise, we train a multi-channel own voice reconstruction system. Experiments using real noise are carried out to investigate the generalization capability. Results show that incorporating augmented noise yields large benefits, in particular considering individualized noise augmentation leads to higher performance.

A.1 Introduction

In noisy working environments such as industrial production or surgery rooms, communication is often impaired. Radio communication based on own voice pickup and transmission can considerably improve communication [1]. In applications where the safety equipment such as breathing masks, protective helmets, or earmuffs prevents using a close-talk microphone in front of the talkers mouth, in-the-ear hearables with integrated microphones are more suited to record the own voice of the talker. Both an outer microphone (OM) and an in-ear microphone (IM) can be beneficial for systems aiming at reconstructing the own voice of the hearable user from noisy recordings. However, OM recordings suffer from external noise, and IM recordings suffer from low-frequency amplification and band-limitation (occlusion), and external and body-produced noise. Therefore, an own voice reconstruction system estimating clean broadband speech from noisy hearable microphone recordings is required to enable communication. Current approaches usually rely on deep learning, where large amounts of training data are required. In addition to relying on the availability of device-specific own voice recordings, external noise has to be accounted for during training. For inward facing body-conduction sensors, it is often assumed no external noise is picked up [2–5]. While other multi-microphone DNN-based systems for hearing device speech enhancement are often trained using artificial head impulse responses for noise, e.g. [6, 7], it has previously not been investigated whether this is sufficient for systems utilizing IMs.

In [3, 4] it has been proposed to reconstruct the own voice from body-conduction sensor recordings using a deep neural network (DNN) without accounting for external noise. In [8], a bandwidth-extension based approach for hearable IM recordings has been trained similarly, but with added body-produced noise recordings during training. In [2, 5], multi-channel systems utilizing both an OM and an inward facing body-conduction sensor are proposed where during training, noise is only added to the OM signal. In [9], a dictionary-based approach has been proposed in which OM and noise at a contact microphone are modeled independently. In [10], a multi-channel system for automatic own voice speech recognition has been proposed which was trained by adding noise only to the OM signals. In [11], real own voice recordings with a bone conduction sensor have been used for training a reconstruction system without accounting for external noise. Fine-tuning with noise recorded from different talkers was investigated. For IMs however, previous research suggests that the transmission of external noise through the device does not only depend on the device used, but also on individual differences [12] and the direction of arrival [13].

In this paper, we propose several techniques to simulate external noise at hearable microphones for use in data augmentation. To our knowledge, this is the first work to address noise data augmentation in the context of own voice reconstruction with an OM and an IM. The influence of noise augmentation techniques is evaluated based on real speech and noise recordings. Results show that the use of the proposed techniques, in particular the use of individualized noise augmentation, leads to superior performance. In an ablation study, we find that in low signal-to-noise ratios

(SNRs) the contribution of the IM is larger, while in high SNRs high performance can be achieved using only the OM.

A.2 Signal model

Fig. A.1 depicts the considered scenario, where a talker is wearing an in-the-ear hearable device equipped with an OM and an IM, denoted by subscript o and i, respectively. In the short-time Fourier transform (STFT) domain, the noisy own voice signal of talker a recorded at the OM is denoted by $Y_o^a(k, l)$ where k is the frequency bin index and l is the time frame index. The recorded noisy own voice signal at the OM consists of an own voice component $S_o^a(k, l)$ and an external noise component $N_o^a(k, l)$, i.e.

$$Y_o^a(k, l) = S_o^a(k, l) + N_o^a(k, l). \quad (\text{A.1})$$

Since the IM is located inside the occluded ear canal, body-produced sounds such as breathing or heartbeats are also recorded [14]. Hence, we define the noisy own voice signal recorded at the IM as

$$Y_i^a(k, l) = S_i^a(k, l) + N_i^a(k, l) + U_i^a(k, l), \quad (\text{A.2})$$

where $U_i^a(k, l)$ denotes body-produced noise. We further assume that the external noise components at the OM and the IM are related by a linear, time-invariant, direction-dependent relative transfer function (RTF) $G_{o,i}^a(k, \theta)$, so that for a single spatially stationary noise source from direction θ the external IM noise component is

$$N_i^a(k, l) = N_o^a(k, l) \cdot G_{o,i}^a(k, \theta). \quad (\text{A.3})$$

For noise fields consisting of several sources, the noise components of each source are assumed to add up to the recorded noise field.

In this work, the goal is to obtain an own voice reconstruction system able to estimate $S_o^a(k, l)$ from the noisy recordings $Y_o^a(k, l)$ and $Y_i^a(k, l)$ obtained from a single hearable device. For training such a system, different methods of simulating external noise transmission during training are investigated.

A.3 Noise data augmentation

Transfer function measurements can be utilized to simulate a large corpus of multi-channel hearable recordings of external noise based on a large single-channel noise corpus. Using a single-channel recording at a reference microphone with STFT $N^{\text{ref}}(k, l)$, the recorded noise at both considered hearable microphones can be modeled using an RTF. For the OM of a hearable worn by talker a , we assume the augmented external noise $\hat{N}_o^a$ is equal to the single-channel recording, so that

$$\hat{N}_o^a(k, l) = N^{\text{ref}}(k, l). \quad (\text{A.4})$$

For the IM, the augmented noise $\hat{N}_i^a$ is modeled by the RTF $\hat{G}_{o,i}$:

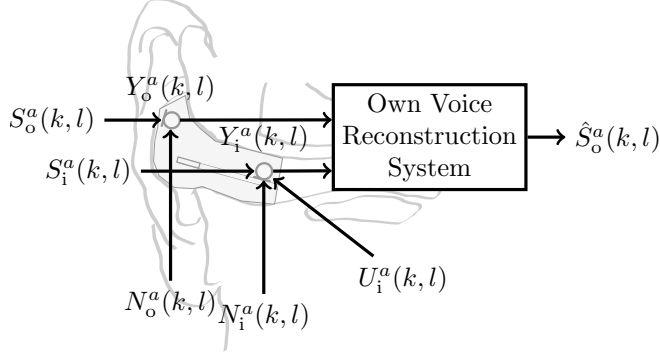


Figure A.1: Noisy multi-microphone own voice reconstruction.

- **No IM noise** Assuming no external noise arrives at the IM, i.e. $\hat{G}_{o,i} = 0$, the noise component is equal to

$$\hat{N}_i^a(k, l) = 0. \quad (\text{A.5})$$

- **Artificial head** Assuming the RTF depends on the direction of arrival, but neglecting individual differences, we propose to use a set of measurements for different directions θ using an artificial head (AH):

$$\hat{N}_i^a(k, l, \theta) = \hat{N}_o^a(k, l) \cdot \hat{G}_{o,i}^{\text{AH}}(k, \theta). \quad (\text{A.6})$$

- **Non-individual** Accounting only for directional differences but instead of an AH using RTFs from a single talker b , we propose to model the noise component as

$$\hat{N}_i^a(k, l, \theta) = \hat{N}_o^a(k, l) \cdot \hat{G}_{o,i}^b(k, \theta). \quad (\text{A.7})$$

- **Individual** Assuming the RTF is subject to individual differences and direction of arrival, both direction-dependent and individual variations can be accounted for by using directional RTFs for each individual talker:

$$\hat{N}_i^a(k, l, \theta) = \hat{N}_o^a(k, l) \cdot \hat{G}_{o,i}^a(k, \theta). \quad (\text{A.8})$$

During speech and noise mixing, a is the same talker an own voice and augmented noise recording to be mixed.

In order to simulate single sources for obtaining multi-channel noise data for DNN training, the direction θ is chosen at random. Since realistic noisy environments often consist of more than one noise source, we also simulate pseudo-diffuse noise with the direction-dependent methods. In order to simulate pseudo-diffuse noise, noise signals from all possible source directions θ are obtained using one of the methods in (A.5)-(A.8) and added:

$$\hat{N}_i^{a, \text{diff}}(k, l) = \sum_{\theta} \hat{N}_i^a(k, l, \theta). \quad (\text{A.9})$$

Each directional reference noise signal is further delayed by one second, so that no two directional signals can be synchronous.

For both pseudo-diffuse or single-source noise, white noise is added to RTF-based IM noise with a random level in $[-\infty, -60]$ dB relative to the IM external noise component. This procedure reduces the coherence between the IM and OM noise signals, bringing it closer to measured coherence of real external noise recordings. For each recording it is randomly decided whether a single source noise or pseudo-diffuse noise is obtained with a probability of 0.5 each during training with each augmentation method.

A.4 Experimental setup

A.4.1 Datasets

For data augmentation, approximately 180 h of single-channel noise recordings from the fifth DNS challenge [15] are used. The AH RTFs are obtained from the Hearpiece database [12] where the closed-vent variant of a prototype hearable [16] is used. Directional RTFs are chosen either as 8 horizontal directions in 45° -steps (Artificial head), or with fine resolution in 7.5° -steps (Artificial head fine). The concha mic. is chosen as the OM. From 18 individual talkers wearing the same hearable, external noise RTFs are measured for 8 horizontal directions in 45° -steps using exponential sweeps from 80 Hz to 22.05 kHz with a duration of 3 s played from 8 loudspeakers in a circle with a distance of 1.5 m around the talker. Both sets of RTFs as well as the single-channel noise recordings are only used for simulating training data. For non-individual methods, a random talker is chosen, and for the non-individual non-directional method, a random direction is chosen as well. From the same talkers, individual multi-channel external noise recordings were obtained in the same loudspeaker configuration. The following noise types were recorded: single-source surgery room noise, metal grinder, directional babble, pseudo-diffuse babble, pseudo-diffuse surgery room noise, pseudo-diffuse factory noise. These real noise recordings are only used for testing. From the same talkers, approximately 25–30 minutes of German own voice speech per talker were recorded in a sound-proofed listening booth while they were wearing the hearable devices. As these recordings are obtained in-situ, the body-produced noises are also recorded at the IM. Recordings are split into train/validation/test parts consisting of 12/2/4 talkers without overlap.

A.4.2 Training details

Audio files are resampled to 16 kHz. Own voice utterances are cut to 3 s length. In this work, only recordings from the left ear device are used. Speech and noise recordings are mixed to a range of $[-10, 25]$ dB SNR defined at the OM, and the IM noise is scaled accordingly so that noise level differences are preserved. Mean-variance normalization is applied to the noisy signal for each microphone individually. The noisy signal statistics of the OM are also utilized to scale the target speech sig-

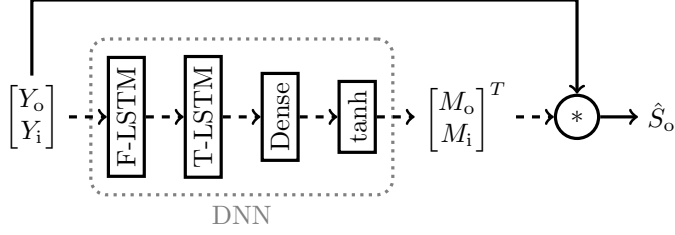


Figure A.2: DNN-based own voice reconstruction system utilizing outer and inner microphone for mask estimation and filtering (OM+IM) based on the FT-JNF architecture from [22].

DNN	Input	Output	Own voice estimate
OM	Y_o	M_o	$\hat{S}_o = M_o \cdot Y_o$
IM	Y_i	M_i	$\hat{S}_o = M_i \cdot Y_i$
OM+auxIM	Y_o, Y_i	M_o	$\hat{S}_o = M_o \cdot Y_o$
OM+IM	Y_o, Y_i	M_o, M_i	$\hat{S}_o = \sum_{m \in \{i,o\}} M_m \cdot Y_m$

Table A.1: DNN variants with different microphone contributions to mask estimation and STFT filtering. Here, auxIM indicates the IM is only used as auxiliary input for mask estimation, but not as a signal to be filtered. STFT and talker indices are omitted for the sake of readability.

nal by the same amount as the speech component in the noisy signal as in [17]. STFTs are computed with frame size of 512 samples corresponding to 32 ms and 50 % overlap, where both in analysis and synthesis a square-root Hann window is used. A batch size of 4 is used in training. We utilize the combined L_1 loss of the time-domain and STFT-domain estimated and target speech signals [18]. The Adam optimizer [19] with learning rate 10^{-4} is used. Training is carried out to a maximum of 100 epochs. The learning rate is halved after 3 consecutive epochs without validation loss improvement and early stopping is applied after 6 consecutive epochs without improvement. For the noise augmentation experiment, the OM+IM DNN variant is trained with real own voice recordings and external noise obtained from using the different augmentation methods described in Section A.3. Testing is carried out using real own voice recordings and real external noise recordings. Trained system performance is evaluated in terms of wideband PESQ [20] and STOI [21]. The clean OM speech signal is used as reference for both metrics. Results are averaged over talkers and noise types.

A.4.3 DNN architecture

We utilize the FT-JNF architecture proposed in [22] (see Fig. A.2) with uni-directional LSTM layers. The architecture follows an STFT-based masking approach, of which we consider several variants with different microphone contributions to mask estimation and STFT masking. The details of each variant are listed in Table A.1. The DNNs compute the complex-valued STFT masks M_o and/or M_i , which are multiplied with the noisy STFTs Y_i and Y_o in a weighted overlap-add scheme. If the DNN is not trained to output the mask M_m for microphone m , the corresponding channel is not used in filtering. The DNN variants differ only in their input and output dimension. The first LSTM has 512, the second has 128 hidden units. The architecture variants consist of around 1.4M parameters each.

A.5 Results

The results of the noise augmentation experiment are presented in Section A.5.1. To investigate the contribution of each channel as auxiliary or filtering input, the results of an ablation study are presented in Section A.5.2.

A.5.1 Noise augmentation

The results of the noise augmentation experiment are shown in Fig. A.3. If no IM noise is considered during training, the trained DNNs improve PESQ scores over the noisy OM signals. For low SNRs, the STOI of the processed signal is higher than of the noisy OM, but for high SNRs, it is lower. Further improvement is gained by using AH RTFs for data augmentation instead of no incorporating IM noise during training. When more directions are considered using fine instead of coarse resolution, the benefit is higher. When non-individual talker RTFs are used instead of an AH, there is further improvement. Although the non-individual method utilizes less RTF measurements than the artificial head method with fine resolution, it achieves slightly higher scores. Finally, if individual RTFs are used in data augmentation, the highest own voice reconstruction performance is achieved. Overall, there is a consistent gain from simulating IM noise during training by using any of the proposed noise augmentation methods.

A.5.2 Microphone contribution ablation study

The results of the ablation study are shown in Fig. A.4. The OM-DNN yields large improvements over the noisy OM signals in high SNRs, but smaller improvements in low SNRs. The performance of OM-DNN is better than IM-DNN for high SNRs, but worse for low SNRs. When both microphone signals are used as input for filtering, the OM+auxIM-DNN improves over the OM-DNN for all SNRs, while only yielding better scores than the IM-DNN above 0 dB SNR. The OM+IM-DNN further improves performance over the OM+auxIM-DNN through the use of the

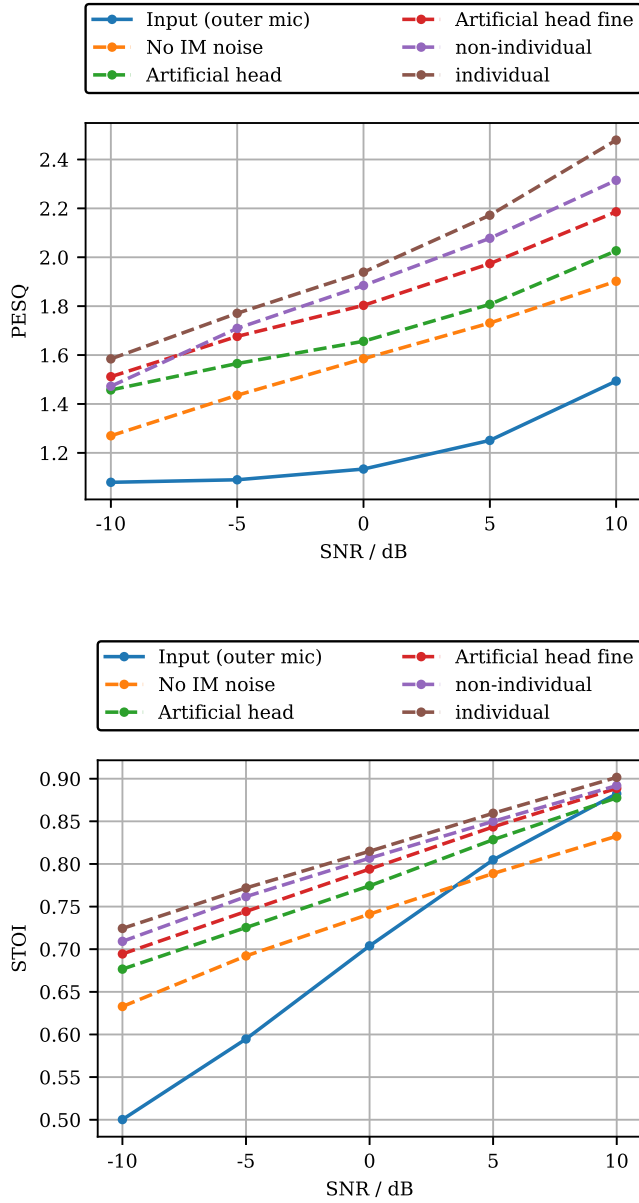


Figure A.3: Results of the noise augmentation experiment. Here, the DNN variant which utilizes both outer and inner microphone in mask estimation and filtering (OM+IM) is used.

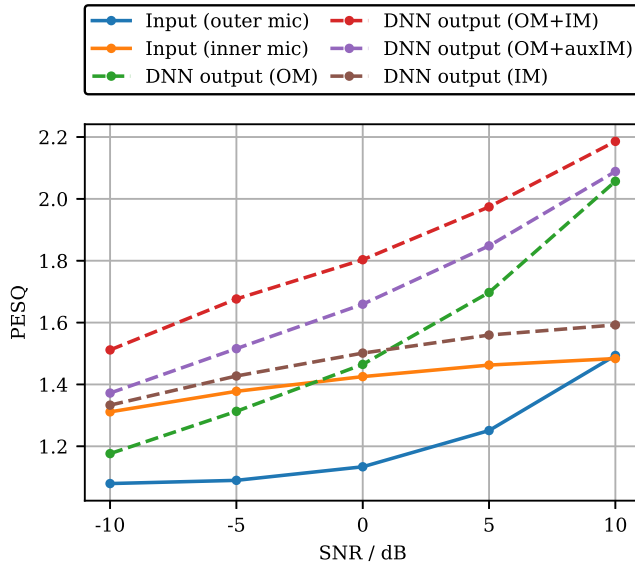


Figure A.4: Channel contribution ablation study results. Augmented noise using AH RTFs with fine resolution was used for training.

IM signals as input in filtering. Overall, we note a large benefit from using the IM in low SNRs, while in high SNRs the contribution of the OM is larger.

A.6 Conclusion

In this paper, we have proposed multi-microphone noise augmentation methods for DNN-based own voice reconstruction. Noise augmentation schemes for training a multi-microphone own voice reconstruction system were evaluated. Experimental results show that incorporating noise augmentation in training of the considered own voice reconstruction system is beneficial. Using individualized noise augmentation leads to the best performance. Additionally, we have investigated the SNR-dependent benefit of an IM, which is high especially in low SNRs.

References

- [1] S. Nordholm, A. Davis, P. C. Yong, and H. H. Dam, “Assistive listening headsets for high noise environments: Protection and communication,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Apr. 2015, pp. 5753–5757.
- [2] H. Wang, X. Zhang, and D. Wang, “Fusing Bone-Conduction and Air-Conduction Sensors for Complex-Domain Speech Enhancement,” *IEEE/ACM*

- Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 3134–3143, 2022.
- [3] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, “EBEN: Extreme bandwidth extension network applied to speech signals captured with noise-resilient body-conduction microphones,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, Jun. 2023.
 - [4] H.-P. Liu, Y. Tsao, and C.-S. Fuh, “Bone-conducted speech enhancement using deep denoising autoencoder,” *Speech Communication*, vol. 104, pp. 106–112, Nov. 2018.
 - [5] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, “Time-Domain Multi-Modal Bone/Air Conducted Speech Enhancement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.
 - [6] M. Tammen and S. Doclo, “Deep Multi-Frame MVDR Filtering for Binaural Noise Reduction,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022.
 - [7] N. L. Westhausen and B. T. Meyer, “Low bit rate binaural link for improved ultra low-latency low-complexity multichannel speech enhancement in Hearing Aids,” in *arXiv*, 2023.
 - [8] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Training Strategies for Own Voice Reconstruction in Hearing Protection Devices Using An In-Ear Microphone,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022.
 - [9] M. Tammen, X. Li, S. Doclo, and L. Theverapperuma, “Dictionary-Based Fusion of Contact and Acoustic Microphones for Wind Noise Reduction,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022.
 - [10] M. Wang, J. Chen, X. Zhang, Z. Huang, and S. Rahardja, “Multi-modal speech enhancement with bone-conducted speech in time domain,” *Applied Acoustics*, vol. 200, p. 109 058, Nov. 2022.
 - [11] Y. Li, Y. Wang, X. Liu, Y. Shi, S. Patel, and S.-F. Shih, “Enabling Real-Time On-Chip Audio Super Resolution for Bone-Conduction Microphones,” *Sensors*, vol. 23, no. 1, p. 35, Jan. 2023.
 - [12] F. Denk and B. Kollmeier, “The Hearpiece database of individual transfer functions of an in-the-ear earpiece for hearing device research,” *Acta Acustica*, vol. 5, 2021.
 - [13] S. Liebich, J.-G. Richter, J. Fabry, C. Durand, J. Fels, and P. Jax, “Direction-of-arrival dependency of active noise cancellation headphones,” in *ASME 2018 Noise Control and Acoustics Division Session presented at INTERNOISE*, Aug. 2018.
 - [14] R. E. Bouserhal, A. Bernier, and J. Voix, “An in-ear speech database in varying conditions of the audio-phonation loop,” *J. Acoust. Soc. Am.*, vol. 145, no. 2, pp. 1069–1077, Feb. 2019.

- [15] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, H. Gamper, M. Golestaneh, and R. Aichner, “Deep Speech Enhancement Challenge at ICASSP 2023,” in *arXiv*, 2023.
- [16] F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J.-H. Bach, J. Wellmann, and B. Kollmeier, “A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research,” in *Proc. AES International Conference on Headphone Technology*, San Francisco, USA, Aug. 2019, pp. 1–9.
- [17] S. Braun and I. Tashev, “Data augmentation and loss normalization for deep noise suppression,” in *Int. Conf. on Speech and Computer (SPECOM)*, vol. 22, St. Petersburg, Russia, Oct. 2020, pp. 79–86.
- [18] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, “STFT-Domain Neural Speech Enhancement with Very Low Algorithmic Latency,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 397–410, 2023.
- [19] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015.
- [20] 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,” *International Telecommunications Union*, Feb. 2001.
- [21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [22] K. Tesch and T. Gerkmann, “Insights Into Deep Non-Linear Filters for Improved Multi-Channel Speech Enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023.

LIST OF PUBLICATIONS

The following publications and datasets are related to the work in this thesis:

Peer-reviewed Journal Papers

- [J3] M. Ohlenbusch, C. Rollwage, S. Doclo, and J. Rennie, “Subjective quality evaluation of personalized own voice reconstruction systems,” *Acta Acustica*, vol. 10, no. 26, 2026. DOI: [10.1051/aacus/2026021](https://doi.org/10.1051/aacus/2026021).
- [J2] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Speech-dependent data augmentation for own voice reconstruction with hearable microphones in noisy environments,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, no. 32, 2025. DOI: [10.1186/s13636-025-00418-1](https://doi.org/10.1186/s13636-025-00418-1).
- [J1] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Modeling of speech-dependent own voice transfer characteristics for hearables with an in-ear microphone,” *Acta Acustica*, vol. 8, no. 28, 2024. DOI: [10.1051/aacus/2024032](https://doi.org/10.1051/aacus/2024032). [Online]. Available: [10.1051/aacus/2024032](https://doi.org/10.1051/aacus/2024032).

Peer-reviewed Conference Papers

- [C6] M. Ohlenbusch, M. Kegler, and M. Stamenovic, “PAS-SE: Personalized auxiliary-sensor speech enhancement for voice pickup in hearables,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2026, pp. 18 942–18 946. DOI: [10.1109/ICASSP55912.2026.11460554](https://doi.org/10.1109/ICASSP55912.2026.11460554).
- [C5] R. Metzger, M. Ohlenbusch, C. Rollwage, and S. Doclo, “Comparison of knowledge distillation methods for low-complexity multi-microphone speech enhancement using the FT-JNF architecture,” in *Proc. ITG Conference on Speech Communication*, Berlin, Germany, 2025, pp. 131–135.
- [C4] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Low-complexity own voice reconstruction for hearables with an in-ear microphone,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, Apr. 2025. DOI: [10.1109/ICASSP49660.2025.10887874](https://doi.org/10.1109/ICASSP49660.2025.10887874).
- [C3] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Multi-microphone noise data augmentation for DNN-based own voice reconstruction for hearables in noisy

environments,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 416–420. DOI: [10.1109/ICASSP48485.2024.10447066](https://doi.org/10.1109/ICASSP48485.2024.10447066).

- [C2] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Speech-dependent modeling of own voice transfer characteristics for in-ear microphones in hearables,” in *Proc. Forum Acusticum*, Turin, Italy, Sep. 2023, pp. 1899–1902. DOI: [10.61782/fa.2023.1030](https://doi.org/10.61782/fa.2023.1030).
- [C1] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Training strategies for own voice reconstruction in hearing protection devices using an in-ear microphone,” in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022. DOI: [10.1109/IWAENC53105.2022.9914801](https://doi.org/10.1109/IWAENC53105.2022.9914801).

Datasets

- [D3] M. Ohlenbusch, C. Rollwage, S. Doclo, and J. RENNIES, *Subjective ratings and objective metric predictions of generic and personalized own voice reconstruction systems*, Zenodo, Apr. 2025. DOI: [10.5281/zenodo.15248719](https://doi.org/10.5281/zenodo.15248719).
- [D2] M. Ohlenbusch, C. Rollwage, and S. Doclo, *Transfer function measurements for simulating environmental noise at hearable microphones*, Zenodo, May 2024. DOI: [10.5281/zenodo.11196867](https://doi.org/10.5281/zenodo.11196867).
- [D1] M. Ohlenbusch, C. Rollwage, and S. Doclo, *German own voice recordings with hearable microphones*, Zenodo, Mar. 2024. DOI: [10.5281/zenodo.10844599](https://doi.org/10.5281/zenodo.10844599).