

Low-Complexity Own Voice Reconstruction for Hearables with an In-Ear Microphone

Mattes Ohlenbusch*, Christian Rollwage*, Simon Doclo*[†]

*Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Germany

[†]Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

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.

Index Terms—own voice reconstruction, hearables, speech enhancement, low-complexity, data augmentation

I. 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.

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.

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.

II. SIGNAL MODEL

We consider a hearable device equipped with an outer microphone and an in-ear microphone, as depicted in Fig. 1. The signals are denoted by subscripts o for the outer microphone and i for the in-ear microphone. In the short-time Fourier

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

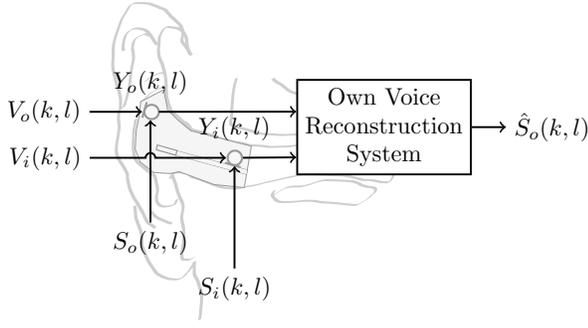


Figure 1. Block diagram of own voice reconstruction using an outer and an in-ear microphone of a hearable.

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

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

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

where the noise components are denoted by $V_o(k, l)$ and $V_i(k, l)$. We assume the noise components 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.

III. 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. 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-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.,

$$\hat{S}_o(k, l) = \sum_{m \in \{o, i\}} M_m(k, l) \cdot Y_m(k, l). \quad (3)$$

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. 2 for the extra-large (XL), large (L),

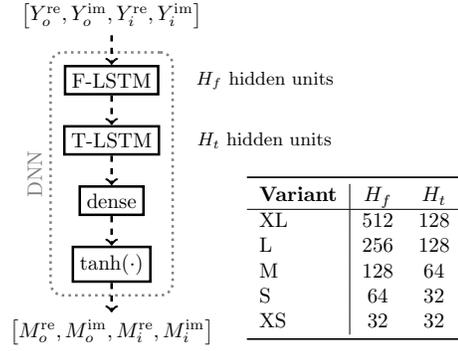


Figure 2. Architecture of the OVR system, estimating complex-valued masks for the outer and in-ear microphones. The number of hidden units, H_f and H_t , is different for each proposed variant.

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].

IV. 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) between 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

$$\hat{S}_i(k, l) = \hat{H}_{p_o(l)}(k) \cdot S_o(k, l). \quad (4)$$

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.

V. EXPERIMENTAL SETUP

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

A. 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 IV. 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 VI-B. 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).

B. 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

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

stopped after six epochs without improvement. The initial learning rate for fine-tuning is 10^{-5} .

C. 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).

D. 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 V-B for the proposed FT-JNF variants:

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

VI. RESULTS

In this section, the results of the experimental evaluation are presented. In Section VI-A, the proposed FT-JNF variants are compared to the baseline systems. In Section VI-B, 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.

A. Comparison to baseline systems

Table I 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

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

Table 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.

System	Intrusive metrics			Size and complexity		
	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

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.

B. 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

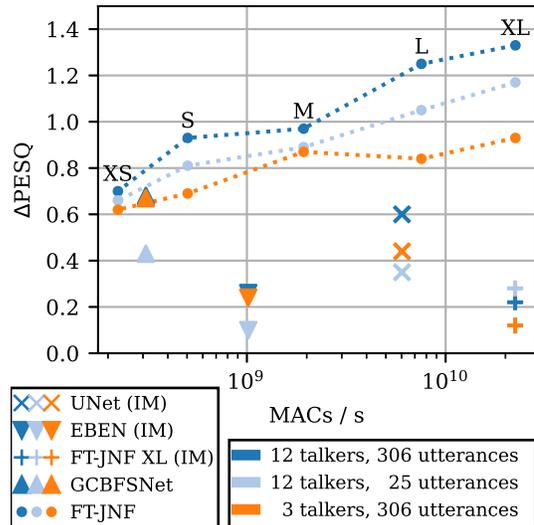


Figure 3. PESQ improvement of the baseline systems and the proposed FT-JNF variants for different amounts of device-specific recorded signals (talkers, utterances). Different systems are distinguished by different symbols, while different amounts of recorded signals are represented by different colors.

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. 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 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.

VII. 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] Rachel E. Bouserhal, Antoine Bernier, and Jérémie 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] Mattes Ohlenbusch, Christian Rollwage, and Simon 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, Sept. 2022.
- [3] Julien Hauret, Thomas Joubaud, Véronique Zimpfer, and Éric 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] Amin Edraki, Wai-Yip Chan, Jesper Jensen, and Daniel 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. 10456–10460.
- [5] Yueyuan Sui, Minghui Zhao, Junxi Xia, Xiaofan Jiang, and Stephen 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] Changtao Li, Feiran Yang, and Jun 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.
- [7] Changtao Li, Feiran Yang, and Jun 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.
- [8] Cheng Yu, Kuo-Hsuan Hung, Syu-Siang Wang, Yu Tsao, and Jeih-Weih Hung, “Time-Domain Multi-Modal Bone/Air Conducted Speech Enhancement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.
- [9] Mou Wang, Junqi Chen, Xiaolei Zhang, Zhiyong Huang, and Susanto Rahardja, “Multi-Modal Speech Enhancement with Bone-Conducted Speech in Time Domain,” *Applied Acoustics*, vol. 200, no. 109058, Nov. 2022.
- [10] Heming Wang, Xueliang Zhang, and DeLiang 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] Mattes Ohlenbusch, Christian Rollwage, and Simon 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.
- [12] Mattes Ohlenbusch, Christian Rollwage, and Simon Doclo, “Speech-dependent Data Augmentation for Own Voice Reconstruction with Hearable Microphones in Noisy Environments,” *arXiv:2405.11592*, May 2024.
- [13] Julien Hauret, Malo Olivier, Thomas Joubaud, Christophe Langrenne, Sarah Poirée, Véronique Zimpfer, and Éric Bavu, “Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors,” *arXiv:2407.11828*, July 2024.
- [14] Wangyou Zhang, Kohei Saijo, Jee weon Jung, Chenda Li, Shinji Watanabe, and Yanmin Qian, “Beyond Performance Plateaus: A Comprehensive Study on Scalability in Speech Enhancement,” in *Proc. Interspeech*, Kos, Greece, Sept. 2024, pp. 1740–1744.
- [15] Kristina Tesch and Timo 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] Marko Stamenovic, Nils L. Westhausen, Li-Chia Yang, Carl Jensen, and Alex 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] Mattes Ohlenbusch, Christian Rollwage, and Simon Doclo, “Modeling of speech-dependent own voice transfer characteristics for hearables with an in-ear microphone,” *Acta Acust.*, vol. 8, 2024.
- [18] Florian Denk, Miriam Lettau, Henning Schepker, Simon Doclo, Reinhild Roden, Matthias Blau, Jörg-Hendrik Bach, Jan Wellmann, and Birger 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] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, “Common Voice: A Massively-Multilingual Speech Corpus,” in *Proc. 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4218–4222.
- [20] Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Alex Ju, Mehdi Zohourian, Min Tang, Mehrsa Golestaneh, and Robert Aichner, “ICASSP 2023 Deep Noise Suppression Challenge,” *IEEE Open Journal of Signal Processing*, vol. 5, pp. 725–737, 2024.
- [21] Zhong-Qiu Wang, Gordon Wichern, Shinji Watanabe, and Jonathan 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] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *Proc. International Conference Learned 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] Jesper Jensen and Cees 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] Augustine Gray and John Markel, “Distance Measures for Speech Processing,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.
- [26] Heming Wang and DeLiang Wang, “Towards Robust Speech Super-Resolution,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 2058–2066, 2021.
- [27] Nils L. Westhausen and Bernd 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.