Enhancing Voice Activity Detection for an Elderly-Centric Self-Learning Conversational Robot Partner in Noisy Environments

Subashkumar Rajanayagam[1], Max Andreas Ingrisch[1], Pascal Müller[2], Patrick Jahn[2] and Stefan Twieg[1]

[1]*Department of Electrical, Mechanical and Industrial Engineering, Anhalt University of Applied Sciences, Bernburger Str. 57, 06366 Koethen, Germany*

[2]*Health Service Research Working Group|Acute Care, Department of Internal Medicine, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Magdeburger Str. 12, 06112 Halle (Saale), Germany*

{subashkumar.rajanayagam, max.ingrisch, stefan.twieg}@hs-anhalt.de, {pascal.mueller, patrick.jahn}@uk-halle.de

Keywords: Voice Activity Detection, Human Robot Interaction, Conversational Robot Partner, Elderly-Centric.

Abstract: Voice Activity Detection (VAD) is a root component in Human-Robot Interaction (HRI), especially for use cases such as a self-learning personalized conversational robot partner designed to support elderly users with high acceptance. While state-of-the-art, lightweight deep-learning–based VAD models achieve high precision, they often struggle with low recall in environments with significant background noise or music. In contrast, traditional lightweight rule-based VAD methods tend to yield higher recall but at the expense of precision. These limitations can negatively affect user experience, particularly among elderly individuals, by causing frustration from missed spoken inputs and reducing overall usability and acceptance of the conversational robot partners. This study investigates noise-suppressing preprocessing techniques to enhance both the recall and precision of existing VAD systems. Experimental results demonstrate that effective noise suppression prior to VAD processing substantially improves voice detection accuracy in noisy settings, ultimately promoting better interaction quality in elderly-centric robotic applications. Moreover, optimal sample rate, frame duration, thresholds and voice activity modes were identified for the robot Double3—the conversational robot partner platform for seniors in a care home, co-creatively developed by reflecting with the nursing staff. An open-source dataset and a dataset collected and annotated in-house with the Double3 robot were evaluated for robustness in benchmarks.

1 INTRODUCTION

Voice Activity Detection (VAD) is a fundamental component in Human-Robot Interaction (HRI), particularly for a self-learning Conversational Robot Partner (CRP) designed for elderly users operating in indoor noisy environments. Robust VAD is essential for enabling these systems to effectively separate spoken content from ambient noise, thereby supporting self-learning personalized processes and maintaining contextual awareness through speaker diarization, speech recognition, and large language model-based text generation. Achieving high recall ensures no relevant speech is missed, while high precision minimizes the downstream tasks' processing of non-speech segments. This balance enhances interaction quality and optimizes resource utilization, energy consumption, and bandwidth efficiency. [1]

State-of-the-art lightweight deep learning VAD Silero [2] and rule-based VAD WebRTC [3] excel in high precision and recall respectively in environments with background noise or music. However, Silero often suffers from low recall, while WebRTC struggles with low precision. This can frustrate the elderly users and ultimately reduce the usability and acceptance of the system. A noise suppression methodology as a preprocessing step to the VADs is investigated to address these shortcomings without sacrificing their strengths. This approach effectively mitigates the limitations inherent to each system, preserving their advantages while enhancing overall performance.

The CRP born from the project EduXBot[1,2] is developed for German senior residents at a nursing home facility in indoor noisy environments by evaluating and reflecting with the nursing staff. Most modern VADs are mainly developed for English and

[1] https://www.hs-anhalt.de/eduxbot/uebersicht.html
[2] https://drks.de/search/de/trial/DRKS00034195

their range of dominant frequencies lies from 1000 Hz to 2000 Hz, whereas the German language is from 125 Hz to 3000 Hz. [4] However, language agnosticism has been studied and results show no significant difference between the accuracy of German and English languages in various SOTA VAD benchmarks. [5] Therefore, this is not considered in this study.

Furthermore, since supervised learning based VADs such as Silero are susceptible to acoustic mismatch in different environments, experiments with the open-source dataset AVA Speech and a custom-recorded and annotated dataset based on the Double3 robot were conducted. [6][7] This is to prove that no separate training would be necessary, which is normally inherent in unsupervised VADs and noise grouping. [8][9][10][11] Finally, the optimal sampling rate, frame duration and VAD Modes/thresholding value were identified for the CRP development platform Double3 in which the VAD was deployed. The Double3 platform demonstrated consistent acceptance among nursing staff, as evidenced by increased System Usability Scale (SUS) scores across three iterations, rising progressively from below 70 to above 75. [12]

2 BACKGROUND

2.1 Base Models

Two lightweight voice activity detection (VAD) models in our experiments are employed: one based on deep learning and the other rule-based. The deep-learning-based model utilized is Silero, a lightweight VAD trained on datasets covering over 6,000 languages. Silero employs Short-Time Fourier Transform (STFT) as features and operates efficiently at a sampling rate of 16 kHz with a 30ms frame duration. Processing each frame requires less than 1ms on a single CPU thread, making it highly suitable for real-time applications [2]. Additionally, the model allows adjustable parameters, including threshold levels and minimum durations for speech and silence, enabling effective customization for various use cases.

In contrast, WebRTC is a rule-based VAD model utilizing Gaussian Mixture Models (GMM). It processes six frequency bands ranging from 80 Hz to 4000 Hz, represented as log-energy values. Optimized for real-time web communication through fixed-point arithmetic operations, WebRTC is highly suitable for deployment on edge devices due to its lightweight design. It supports multiple sampling rates of 8, 16, and 32 kHz and accepts frame durations of 10, 20, and 30ms. The presence of voice activity is determined by applying predefined rule-based criteria, enhancing its efficiency in real-time scenarios [3].

2.2 Datasets

Our experiments were conducted using two distinct datasets. The first dataset is the open-source AVA Speech dataset, which consists of densely annotated audio clips recorded at sampling rates of 44.1 kHz and 48 kHz. The dataset provides approximately 40 hours of audio segmented into four categories: no speech, clean speech, speech with music, and speech with noise. Specifically, clean speech segments account for 14.55%, speech with music 13.46%, speech with noise 24.32%, and no speech 47.68% of the total duration. The signal-to-noise ratios (SNR) for clean speech, speech with music, and speech with noise segments are 40.8 dB, 11.7 dB, and 16.2 dB, respectively [7].

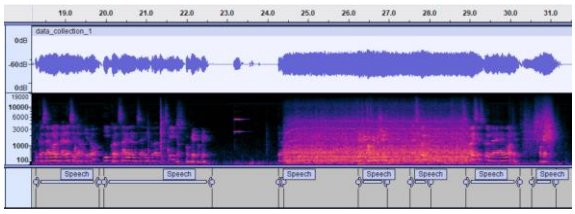

Figure 1: Annotation procedure with Audacity for the in-house Double3 dataset for the first 31.5 seconds.

The second dataset was collected in-house using a Double3 robot's microphone, incorporating typical environmental background noises such as coffee machines, doors opening and closing, keyboard typing, mouse clicks, and background music in an indoor environment similar to the nursing home's communal space. This verification dataset, recorded at a sampling rate of 44.1 kHz, comprises approximately 20 minutes of densely annotated audio. Annotations were conducted manually using the Audacity software [13], by carefully listening, analyzing waveforms and spectrograms (see Figure 1). The speech segments were directly labeled using Audacity, while the remaining non-speech segments were identified and annotated using a Python script based on gaps between speech segments. The speech segment accounts for 58.02% and non-speech 41.98% of the total duration.

2.3 Noise Suppression

Recent advancements in lightweight, real-time deep learning-based noise suppression and speech enhancement methods, such as DeepFilterNet, have significantly improved the feasibility of deploying efficient models on edge devices, including robotic platforms equipped with embedded systems [14]. In our study, the DeepFilterNet2 (DFN2) model was explored as a preprocessing step prior to the VAD stage. [14] DFN2 leverages the harmonic structure inherent in speech signals, achieving efficient speech enhancement with a real-time factor of 0.04, thus ensuring suitability for real-time applications. [14]

3 METHODOLOGY

The first objective of this study is to evaluate the performance and limitations of the lightweight VAD models Silero (deep-learning-based) and WebRTC (GMM-based) by benchmarking them against the AVA Speech and in-house Double3 datasets. Initially, WebRTC was benchmarked against both datasets across various sampling rates (8, 16, and 32 kHz), frame durations (10, 20, and 30ms), and VAD operational modes (0, 1, 2, and 3). The VAD mode "0" corresponds to the lowest detection threshold, while mode "3" represents the highest detection threshold, with incremental steps (1 and 2) between each mode. This resulted in 36 benchmark combinations for each dataset, measuring recall and precision. The benchmarks were calculated based on the total duration of speech detected rather than the number of segments, ensuring better accuracy.

The second objective is to investigate the impact of employing the DeepFilterNet2 (DFN2) noise suppression preprocessing technique to mitigate the limitations identified in the lightweight VAD models while preserving their performance. The final objective is to determine the optimal sampling rate, frame duration, threshold values, and operational modes for the VAD models when applied to the Double3-based CRP.

For Silero, only the threshold parameter was varied, evaluated at equal intervals from 0 to 1, while maintaining fixed sampling rate and frame duration (16 kHz and 30ms, respectively), as these are the only supported settings relevant for this study. The evaluation process was subsequently repeated after applying the DFN2 preprocessing step to assess its impact on Silero's performance.

4 RESULTS

This section follows the structure outlined in the previous methodology section. First, the baseline results obtained using WebRTC and Silero are presented. Subsequently, the results achieved after incorporating the DNF2 noise suppression technique as a preprocessing step are discussed.

4.1 Base Results

WebRTC was benchmarked using the AVA Speech dataset and our in-house Double3 dataset across various sampling rates, frame durations, and VAD modes.

Silero, on the other hand, was benchmarked exclusively at a sampling rate of 16 kHz and a frame duration of 30ms. However, various thresholds ranging from 0 to 1 were explored to assess their impact on performance.

4.1.1 WebRTC without Preprocessing

When looking across the sampling rate axis (x-axis) in Figure 2, no significant changes can be seen in the color and size representing the precision and recall for AVA Speech. In Figure 3, this is seen clearly, where the three colors representing the different sample rates overlay each other in most cases.

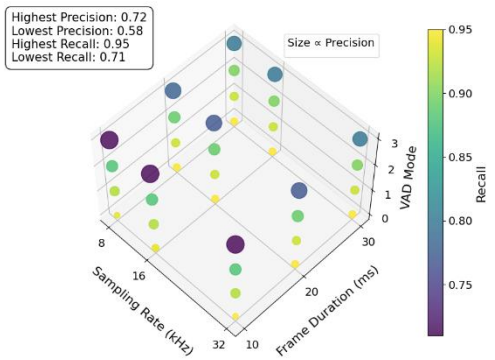

Figure 2: 5D plot of precision vs recall for WebRTC on AVA Speech dataset with various sampling and frame rates without noise suppression preprocessing.

A similar behavior is observed concerning frame duration, although predominantly at lower VAD modes (0 to 2), where threshold values are lower and less aggressive. At higher thresholds—specifically, VAD mode = 3 – the precision peaks at 72% with a frame duration of 10ms; however, this configuration also results in the lowest recall of only 70.8%.

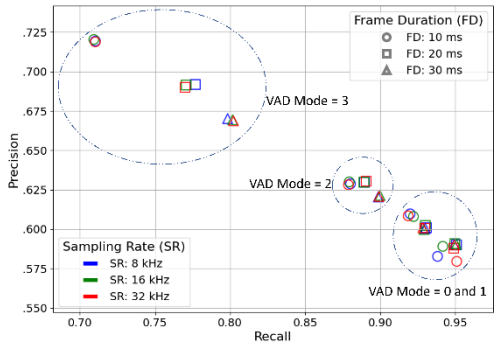

Figure 3: 4D plot of precision vs recall for WebRTC on AVA Speech dataset with various sampling and frame rates without noise suppression preprocessing. VAD modes are bounded by circles.

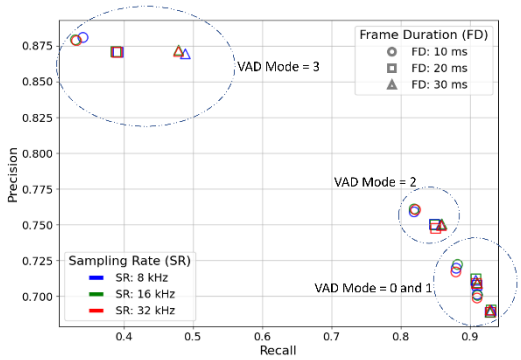

Figure 4: 4D plot of precision vs recall for WebRTC on Double3 in-house dataset for various sampling and frame rates without noise suppression preprocessing. VAD modes bounded by circles.

However, the VAD mode demonstrates a clear relationship between precision and recall for the AVA Speech dataset. As the VAD mode increases, precision improves, whereas recall decreases. At the highest VAD mode (mode 3), precision peaks at 72%, but recall drops to its lowest at 70.8%. Conversely, at the lowest VAD mode (mode 0), precision decreases to the lowest 58%, while recall reaches its highest value of 95%.

Similar observations were made when benchmarking WebRTC on the in-house Double3 dataset, as illustrated in Figure 4. Specifically, the highest precision achieved was 88%, corresponding to the lowest recall value of 33%. Conversely, the highest recall of 93% was associated with the lowest precision of 69%.

Although WebRTC demonstrates high recall for both datasets, peaking above 90%, the precision remains comparatively lower, peaking at 72% for the AVA Speech dataset. These trade-off highlights an inherent limitation of WebRTC VAD, wherein precision and recall cannot be simultaneously optimized, necessitating a compromise between the two metrics.

4.1.2 Silero without Preprocessing

The benchmark results for Silero, illustrating precision versus recall for both the AVA Speech and Double3 datasets, are shown in Figure 5. While Silero demonstrates high precision, exceeding 90% in certain thresholds, its recall remains comparatively lower, peaking at 73% for the AVA Speech dataset and 62% for the Double3 dataset when precision is 90%. Consequently, an optimal balance between precision and recall is not observed in either dataset.

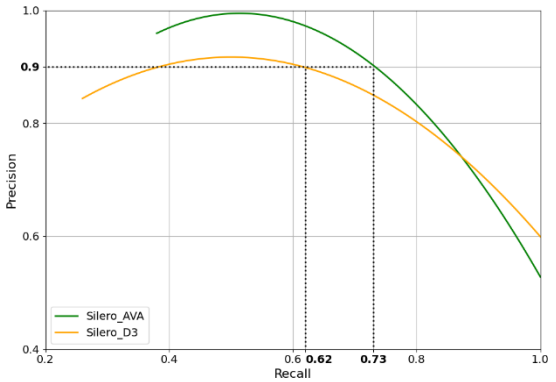

Figure 5: 2D plot of precision vs recall for Silero on AVA Speech (AVA) and Double3 (D3) in-house datasets for various thresholds without noise suppression preprocessing.

In contrast to WebRTC, which demonstrates higher recall but lower precision, Silero exhibits lower recall yet higher precision. However, both systems inherently involve a trade-off between these metrics, which constrains the ability to optimize them simultaneously.

4.2 After Noise Suppression

In this section, the results obtained after applying noise suppression (DNF2) as a preprocessing step to the VAD models WebRTC and Silero are summarized.

4.2.1 WebRTC after DNF2

Benchmarking WebRTC with DNF2 as a noise suppression preprocessing step yielded major improvements in precision. As illustrated in Figures 6 and 7, incorporating the noise suppression

model (DNF2) as a preprocessing step significantly improved the precision of the WebRTC VAD model on both the AVA Speech and Double3 datasets. Specifically, for the AVA Speech dataset, the highest precision increased by 15.5%, rising from 72% to 87.5%. Similarly, the Double3 dataset exhibited a notable improvement, with the highest precision increasing by 5% from 76% to 81%, at VAD mode 2.

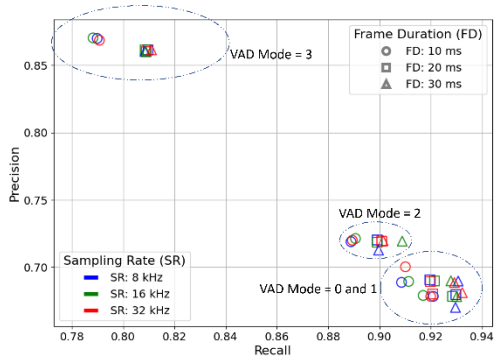

Figure 6: 4D plot of precision vs recall for WebRTC on AVA Speech dataset for various sampling and frame rates after noise suppression. VAD modes bounded by circles.

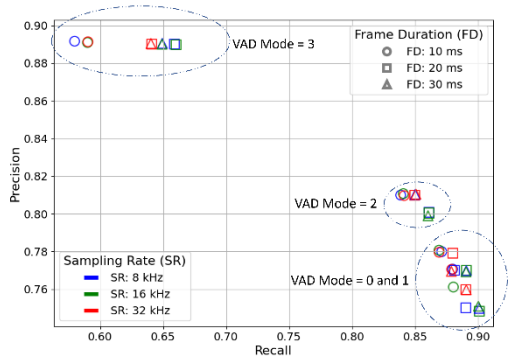

Figure 7: 4D plot of precision vs recall for WebRTC on Double3 dataset for various sampling and frame rates after noise suppression. VAD modes bounded by circles.

The sampling rate showed minimal influence on performance, consistent with observations from the baseline model without noise suppression. However, the frame duration of 10ms in the AVA Speech dataset at VAD Mode 3 achieves recall less than 80%, whereas at frame durations 20ms and 30ms are above 80%. Consequently, an 8 kHz sampling rate combined with a 20ms frame duration was selected to minimize sample size, inference time, and resource usage and maintain a precision and recall above 80% for both datasets.

With this configuration, WebRTC combined with the noise suppression preprocessing step achieved the desired balance between precision and recall. Specifically, at VAD mode 3 for the AVA Speech dataset and VAD mode 2 for the Double3 dataset, both precision and recall exceeded 80%, as seen in Figure 8. Specifically, at this configuration, the AVA Speech dataset precision increased by 17% (from 69% to 86%) and the Double3 dataset precision increased by 5% (from 75% to 80%). In both instances, recall was maintained (Double3 at 85%) or increased (AVA Speech by 3%).

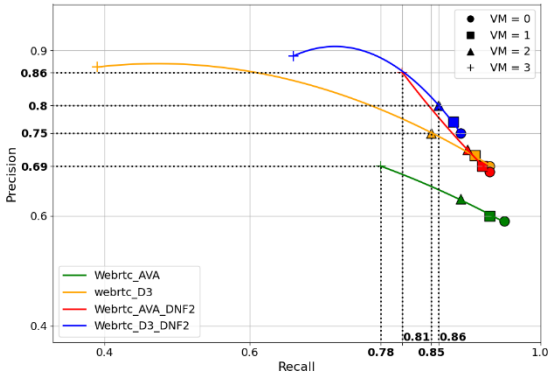

Figure 8: 2D plot of precision vs recall for WebRTC on AVA Speech and Double3 in-house datasets after noise suppression for the selected 8 kHz sampling rate and 20ms frame duration with various thresholds.

4.2.2 Silero after DNF2

Benchmarking Silero with DNF2 as a noise suppression preprocessing step yielded notable improvements in recall.

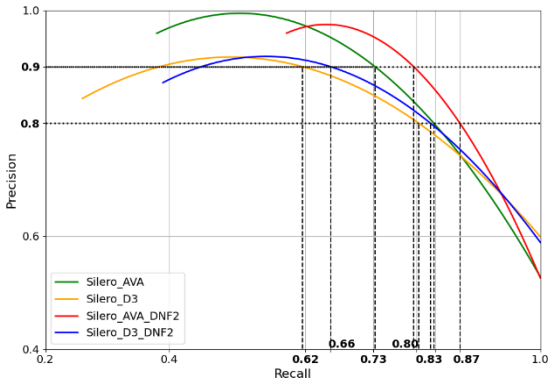

Figure 9: 2D plot of precision vs recall for Silero on AVA Speech and Double3 in-house datasets with various thresholds after noise suppression.

For the AVA Speech dataset, recall increased by 7%, rising from 73% to 80%, at a precision of 90% with a threshold of 0.20, as shown in Figure 9. Similarly, for the Double3 dataset, recall improved by 4% (from 62% to 66%) at the same precision level at a threshold of 0.25.

Although these enhancements are promising, recall remains below 80% in most instances. To explore a more optimal operating point, configuring the system to achieve 80% precision resulted in a recall increase of 4% (from 83% to 87%) for AVA Speech, while for Double3, it rose by 3% (from 80% to 83%) at the thresholds of 0.20 and 0.25 respectively.

5 CONCLUSIONS

In this study, two lightweight VAD models enhanced by incorporating noise suppression (DNF2) as a preprocessing step are evaluated. The addition of this preprocessing significantly improved WebRTC VAD performance, increasing precision by 17% on the AVA Speech dataset and by 5% on the Double3 dataset. This enhancement enabled WebRTC to achieve an optimal combination of precision and recall, with both metrics surpassing 80% at VAD mode 3 for the AVA Speech dataset and mode 2 for the Double3 dataset. The optimal sampling rate of 8 kHz, 20ms frame duration and VAD mode 2 is identified for WebRTC, aligning to the Double3 dataset. Additionally, the deep learning-based Silero model demonstrated improved recall, increasing by 4% for the AVA Speech dataset and 3% for the Double3 dataset, all when precision is at 80%.

Since WebRTC achieved better recall (85%) as compared to Silero (83%) when precision is above 80% for the Double3 dataset, WebRTC with DNF2 noise suppression is selected for the CRP at VAD mode 2.

During the analysis of our results, it was observed that speech characterized by whispering, shouting, and high-pitched tones was frequently missed by both VAD models. This was done by analyzing the before and after spectrograms and listening to the cropped regions of the false negative regions of the audio datasets. Consequently, further optimization and additional benchmarking efforts are required to enhance the detection capabilities and overall performance of these lightweight VAD models.

ACKNOWLEDGMENTS

We gratefully acknowledge the support provided by the Research, Transfer, and Start-Up Center (Forschungs-, Transfer- und Gründerzentrum), Anhalt University of Applied Sciences, and the state of Saxony-Anhalt. We also extend our sincere gratitude to the Federal Ministry of Education and Research (BMBF) for their generous support under grant number 03WIR3118A, which significantly contributed to the successful completion of this study.

REFERENCES

[1] S. Yadav, P. A. D. Legaspi, M. S. O. Alink, A. B. J. Kokkeler and B. Nauta, "Hardware Implementations for Voice Activity Detection: Trends, Challenges and Outlook," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 70, no. 3, pp. 1083-1096, March 2023, doi: 10.1109/TCSI.2022.3225717.

[2] Silero Team, "Silero Models: State-of-the-Art Speech Processing Models," GitHub repository, 2024 [Online]. Available: https://github.com/snakers4/ silero-models. [Accessed: 09-Feb-2025].

Google. (2011) WebRTC [Online]. Available:https://webrtc.org/ [accessed Feb 2025].

[3] Atlantis-vzw.com, ""Learning a foreign language with more ease"." Accessed: Feb. 02, 2025. [Online]. Available: https://www.atlantis-vzw.com/vreemde-talen?lang=en.

[4] M. Sharma, S. Joshi, T. Chatterjee, and R. Hamid, "A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows," Neurocomputing, vol. 494, pp. 116–131, Jul. 2022, doi: 10.1016/J.NEUCOM.2022.04.084.

[5] R. M. Patil and C. M. Patil, "Unveiling the State-of-the-Art: A Comprehensive Survey on Voice Activity Detection Techniques", doi: 10.1109/APCIT62007.2024.10673721.

[6] S. Chaudhuri, J. Roth, D. P. Ellis, A. C. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. Reale, L. G. Reid, K. W. Wilson, and Z. Xi, "AVA-Speech: A densely labeled dataset of speech activity in movies," Aug. 2018. [Online]. Available: https://arxiv.org/abs/1808.00606.

[7] X. L. Zhang and M. Xu, "AUC optimization for deep learning-based voice activity detection," Eurasip J. Audio, Speech, Music Process., vol. 2022, no. 1, pp. 1–12, Dec. 2022, doi: 10.1186/S13636-022-00260-9/TABLES/7.

[8] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-End Automatic Speech Recognition Integrated with CTC-Based Voice Activity Detection," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 6999–7003, May 2020, doi: 10.1109/ICASSP40776.2020.9054358.

[9] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, “A Survey on Deep Learning for Human Activity Recognition,” 2021. A Surv. Deep Learn. Hum. Act. Recognition. ACM Comput. Surv, vol. 54, no. 8, p. 177, 2021, doi: 10.1145/3472290.

[10] S. Twieg and B. Zimmermann. Acoustic clustering for vehicle based sounds, 2010.

[11] P. Müller, H. K. Gali, S. Rajanayagam, S. Twieg, P. Jahn, and S. Hofstetter. Making Complex Technologies Accessible Through Simple Controllability: Initial Results of a Feasibility Study. Applied Sciences, 15(2), 1002, 2024 [Online]. Available: https://doi.org/10.3390/app15021002.

[12] Audacity(R) software is copyright (c) 1999-2014 Audacity Team. [Website: http://audacity.sourceforge.net/. It is free software distributed under the terms of the GNU General Public License.] The name Audacity(R) is a registered trademark of Dominic Mazzoni.

[13] H. Schröter, T. Rosenkranz, A. Escalante-B., and A. Maier, "DeepFilterNet: Perceptually motivated real-time speech enhancement," in Proc. 17th Int. Workshop Acoustic Signal Enhancement (IWAENC), 2022.