Acta Acustica 2022, **6**, 40

https://doi.org/10.1051/aacus/2022035

Available online at:
https://acta-acustica.edpsciences.org

SCIENTIFIC ARTICLE OPEN ACCESS

Near-ear sound pressure level distribution in everyday life considering the user's own voice and privacy

Jule Pohlhausen*, Inga Holube, and Joerg Bitzer

Jade University of Applied Sciences, Institute of Hearing Technology and Audiology, and Cluster of Excellence "Hearing4all", 26121 Oldenburg, Germany

Received 6 October 2021, Accepted 22 August 2022

Abstract – Recently, exploring acoustic conditions of people in their everyday environments has drawn a lot of attention. One of the most important and disturbing sound sources is the test participant's own voice. This contribution proposes an algorithm to determine the own-voice audio segments (OVS) for blocks of 125 ms and a method for measuring sound pressure levels (SPL) without violating privacy laws. The own voice detection (OVD) algorithm here developed is based on a machine learning algorithm and a set of acoustic features that do not allow for speech reconstruction. A manually labeled real-world recording of one full day showed reliable and robust detection results. Moreover, the OVD algorithm was applied to 13 near-ear recordings of hearing-impaired participants in an ecological momentary assessment (EMA) study. The analysis shows that the grand mean percentage of predicted OVS during one day was approx. 10% which corresponds well to other published data. These OVS had a small impact on the median SPL over all data. However, for short analysis intervals, significant differences up to 30 dB occurred in the measured SPL, depending on the proportion of OVS and the SPL of the background noise.

Keywords: Ecological momentary assessment, Own voice detection, Speaking time, Acoustic measurement, Noise dosimeter

1 Introduction

Acoustic measures in everyday life are of great interest for many research questions, such as noise dosimetry or the behavioral analysis of people with hearing impairments. To capture the actual sound exposure of individuals, the microphone(s) should be mounted as close as possible to the ears of the participants, i.e., near-ear. However, since the people's own voice is often the loudest signal at their ear position, analysis results for environmental sounds are influenced by time segments containing the voice of the participant (own voice segments, OVS) [1]. Nevertheless, OVS alone can be of interest for many research fields, e.g., psychological research for dementia or autism [2], or communication situations in hearing aid research. This paper contributes to an application of ecological momentary assessment (EMA) [3] for studying the impact of hearing impairments and hearing aids on the communication abilities of elderly people in natural environments. An important aspect of measurements in natural environments with naive participants is an appropriate recording device that meets relevant laws. In many countries, it is necessary to have written consent of all communication partners (intended or non-intended) when recording audio, which in real-life situations often cannot be obtained. Hence, for long-term monitoring, recordings should be limited to privacy-preserving acoustic features. Therefore, this contribution describes and evaluates an OVS detection (OVD) algorithm based on acoustic features, and it analyses the influence of OVS on long-term monitoring of sound pressure level (SPL) measurements.

Currently, only little data is available on the acoustical characteristics of natural environments, and the duration of stay in those environments for people of different ages, hearing abilities, and life styles. Jensen and Nielsen [4] asked participants with hearing impairments to record audio snippets in their natural environments. They found large SPL variations in all reported categories, with the highest mean of approx. 71 dB SPL for the category *conversations with several persons*, and of approx. 80 dB SPL for mobility locations (car, bus). Wagener et al. [5] also collected snippets of audio recordings from hearing-impaired participants. Their results support the findings of large SPL variations, with a mean SPL generally between 60 dB and 70 dB, and the highest SPL for mobility locations – more than 80 dB. The recordings of communication situations were further analyzed by Smeds et al. [6]. When selecting several seconds of time intervals containing either only background noise or

*Corresponding author: jule.pohlhausen@jade-hs.de

speech plus noise, the calculated signal-to-noise ratio (SNR) was mostly positive, with the highest values for quiet situations and the lowest values for recordings in a moving car.

It is, however, still a crucial question whether valid and useful results should include whole days rather than snapshots of short duration. A suitable sampling strategy would depend on the particular research question. For the purposes of this contribution, continuous measurements are needed to establish the distribution and demands of the situations that are being attended to. Continuous measurements with noise dosimeters were reported by Wu and Bentler [7], who focused on auditory lifestyles. They found lower SPL for older adults than for younger ones, with a mean SPL of 60–73 dB in speech-related events. The same research group confirmed the findings of [6] from audio recordings of hearing-impaired listeners with one microphone at chest height, namely that situations with negative SNR are rarely listened to, that speech SPL increases, and SNR decreases with increasing noise SPL [8]. To date, the largest privacy-preserving data collection in natural environments for mostly elderly hearing-aid users is probably that available from Christensen et al. [9]. They recorded SPL, SNR, and sound modulations, and classified acoustic situations analyzed from hearing aids worn in natural environments throughout the day. The grand median SPL was 54 dB, with lower values at night and higher values at lunch and dinnertime. The highest SPL were measured for the situation *speech in noise*, with a median of 72 dB. *Noise* and *quiet* situations showed median SPL of 66 dB and 49 dB, respectively.

In these studies, SPL were measured either with noise dosimeters or hearing aids, or by analyzing audio recordings. Jensen and Nielsen [4] noted that the SPL measured in communication situations involving one or several persons are affected by the test participant's own voice, probably resulting in small SPL differences between different listening categories. The voice is raised against increasing SPL of the background noise, which is known as the Lombard effect, and has an impact on noise dosimeter measurements [1].

Titze et al. [10] investigated voicing times of 31 teachers over two weeks. They reported that on average voicing occurred for 23% per hour while working, and during leisure time for 12%. The voicing percentage for teachers is consistent with Masuda et al. [11], who reported voicing times of 21% in an eight-hour workday. Mehta et al. [12] measured voicing times in adults (salespersons, clinicians, teachers) for at least five workdays and found no significant difference between normal and disordered voices (both on average 16% per hour). For their so-called low-voice-use group, consisting of laboratory research staff, they reported mean voicing times of only 9% per hour. In Mehl et al. [13], 52 students recorded 30 s audio intervals every 12 min over 2–4 days and the proportion of intervals containing OVS was measured to be approx. 26% on average. In contrast, Fhärm et al. [14] reported a voicing time of 9.9% averaged over 3 eight-hour days for 20 elderly retirees, with a significant difference between women and men. For office workers, Masuda et al. [11] reported a voicing proportion of only 7% in an eight-hour workday. Overall, the voicing times found in the literature cover a wide range and depend on the population studied and the data sampling. All studies reported a large inter-subject variability.

The impact of own voice on measured SPL can be avoided by manually categorizing the audio recordings as "own voice", "other voices", and "background sounds" [6, 8]. However, this approach is time- as well as labor consuming and requires permission to record audio signals in natural environments. This contribution proposes a method for measuring SPL in natural environments that distinguishes between participants' own voice and other sounds, and does not require audio recordings. The developed OVD algorithm is based on machine learning (ML). The work comprises two main research complexes. First, Section 2 describes the OVD algorithm and the evaluation of its performance for manually labeled data. Second, Section 3 shows the predicted proportion of OVS in a near-ear, long-term (NELT) study by von Gablenz et al. [15] with 13 elderly, hearing-impaired participants. Moreover, for the NELT study the distribution of SPL in natural environments was investigated and the difference due to exclusion of the predicted OVS is shown.

2 Own voice segments detection

For the detection of OVS, several algorithms and methods are known. The parameters and implementations vary a great deal, depending on the application. For example, Carullo et al. [16] presented a comparison of methods for clinical voice analysis and voice-dosimeter data recorded over a long time, which use accelerometers placed on the neck of the patient. Hearing aids use the information of OVD to control their signal-processing algorithms [17]. The detection algorithms for monaural systems are manifold and include, e.g., a training phase of the voice [18], classification of the incoming signal [19], or an analysis of the signal in the ear canal by using an additional microphone inside the hearing aid's earmould facing towards the ear drum [20].

This contribution focuses on binaural measures and on analysis procedures without real-time constraints. In comparison to monaural measures binaural measures contain spatial information. For a given setup of two near-ear microphones, applied at equal distance from the participant's mouth, it is assumed that the participant's own voice appears at equal amplitude and in phase at the microphones, while environmental signals usually do not [21]. Previous work of Rasmussen et al. [22] proposed an OVD for hearing aids based on the analysis of the lag in the cross-correlation, whereas Granqvist [21] applied the self-to-other ratio (SOR) as a phonation detection stage for noise dosimetry. Bitzer et al. [23] compared these two methods with an earlier version of the proposed algorithm, and the results showed that the algorithms are not very robust in terms of their detection quality. Therefore, an advanced algorithm using ML is proposed, providing an improved analysis for real-world data. This section specifies the privacy-preserving features, the OVD algorithm, and the labeled data used

for training. Subsequently, the evaluation of the new OVD algorithm using real-world data is described.

2.1 Privacy-preserving features

Our privacy-preserving EMA system olMEGA[1] [24–26] only stored acoustical features (root mean square, RMS; auto and cross-power spectral density, PSD; zero crossing rate, ZCR). All features were calculated online from recordings by two head-worn, near-ear microphones attached to glasses. The system has a flat frequency response within 1 dB between 100 Hz and 3 kHz. Beyond 3 kHz, a slight rise was measured up to 3 dB at 8 kHz [24]. The system was calibrated to SPL at 1 kHz. RMS and ZCR values were computed for 25 ms half-overlapping intervals. The PSD of the 25 ms intervals (samplingrate = 80 Hz) were smoothed with a first-order, low-pass filter with time constants of $\tau = 125$ ms and the smoothed spectra were downsampled to 8 Hz to prevent reconstruction of the original signal. These processing steps led to unrecognizable speech in the reconstructed signal, but the presence of voices and their gender was not hidden [24].

2.2 Training data

The data consist of communication situations recorded with the olMEGA system in the lab and in everyday life, as well as an artificial communication situation in the Gesture Lab at Oldenburg University. In contrast to the intended usage of the olMEGA system and its application in [15], acoustical features and the audio signal were stored with the consent of all persons speaking. The parallel recordings enabled the labeling of the data to obtain reliable training data.

The recordings in the lab were conducted in a lecture room with some acoustical optimizations (acoustic ceiling and curtains, moveable damping items) at Jade University of Applied Sciences in Oldenburg. The room had a volume of 262 m^3 with with a T_{30} of 0.46 s. The conversations of eight different pairs of students took place in a circle of eight loudspeakers used as sources for the background noise. One participant in the center of the loudspeaker circle was wearing the olMEGA system, and was solving a "spot the difference" task (a so-called DiapixUK, [27]) with a communication partner sitting in front of them. The recordings were obtained with five different SPL of background noise, simulating a cafeteria. Each conversation took approx. 5 min, resulting in a total recording duration of 3.79 h. During the lab recordings, most of the time at least one participant spoke. This resulted in a proportion of 77.7% speech (42.8% own voice) and only 22.3% background noise, which is not a realistic distribution.

The recordings in everyday life contain diverse sound scenarios, e.g., in a cafe, in a canteen, in a quiet office, while walking outside, or while driving a car. Overall, the 16 recordings resulted in a total duration of 1.98 h. Compared to the lab recordings, the recordings in everyday life included longer speech pauses, and the proportion of OVS was reduced to 17.8% (total speech 45.2%). All recorded audio signals were manually labeled for speech sources (own voice and other voices).

In the Gesture Lab at the University of Oldenburg, the toolbox for dynamic virtual acoustic environments (TASCAR) [28] was used to simulate a train station, an active street, and a cafeteria with a T_{60} of 1.77 s, 0.14 s, and 1.41 s, respectively [29, 30]. The Gesture Lab consists of 28 loudspeakers, divided into three horizontal parallel circles and four sub-woofers. A G.R.A.S. Head and Torso simulator Type 45BM (also known as KEMAR) wearing the olMEGA system was placed at the center of the circles at a distance of 1 m in front of a NTi Audio TalkBox, to simulate own voice and a communication partner, respectively. Clean speech stereo DiapixUK [27] recordings were used as the two speech sources. Each channel consisted of only one speaker. A simple energy-based detector was applied for labeling speech sequences. Both sources were calibrated to a level of 54 dB SPL at a distance of 0.5 m. Speech presentations and virtual acoustic environments were recorded separately, to allow for subsequent SNR adjustments to 5 dB, 10 dB, and 15 dB SNR relative to the level of the communication partner (NTi Audio TalkBox). In total, 14 conversations were recorded. First, every speaker was used as the own voice, and in a second setup as the voice of the communication partner, resulting in 1.83 h of speech material. Overall, the training data consisted of 7.61 h labeled recordings. The distribution of this data set was balanced between male and female own voice speakers.

2.3 Description of the algorithm

Bitzer and Kissner [31] showed that a smoothed version of the real part of the coherence in the frequency band between 400 Hz and 1 kHz is a relevant indicator of OVS. This measure is always higher for OVS compared to speech from a communication partner, independent of the SNR. It decreases, however, with decreasing SNR [23]. Therefore, a simple threshold, as described in [31], would not yield reliable results in many situations, and an adaptive solution was required.

Therefore, ML was applied and additional features were included to improve robustness and broad applicability of the OVD. The new feature set included the RMS and ZCR, as well as the band energy in 12 octave bands between 62.5 Hz and 12 kHz, derived from the PSD features. Nevertheless, the prediction was limited by the sampling rate of the PSD, i.e., 8 Hz. To capture the behavior in the time domain, the ten-times higher sampling rate of the RMS was exploited and the quartile coefficient of dispersion was calculated

$$\mathrm{QD} = \frac{r_{0.75} - r_{0.25}}{r_{0.5}}, \tag{1}$$

where r_x is the xth-percentile of ten adjacent RMS values (corresponding to 125 ms). The QD provided a measure of

[1] Construction manuals, software implementation, and supplemental Material for olMEGA are available at https://github.com/ol-MEGA (Open Source License for all parts, including hardware design).

stationarity that is robust against outliers. This measure is useful, since background noise has less variation over time than speech [32]. In total, 21 feature values per time segment were extracted.

Although only conversations of younger adults (20–28 years) were recorded as training data, in Section 3 the OVD is applied to NELT-recordings of elderly participants. Since the conversational speech SPL appears to remain stable with increased age [33], the acoustic voice features that are important for OVD with olMEGA as the coherence are less affected by age-related factors and more by inter-individual variations.

The amount of labeled training data was still rather small, and therefore the choice of the ML algorithm was crucial. The Random Forest (RF) algorithm [34] was chosen to binarily classify OVS to all other segments. RF is a supervised ML technique that is robust to noise and has been successfully applied in a wide range of research projects [35]. The current implementation uses the Matlab class TreeBagger included in the "Statistics and Machine Learning Toolbox". The RF algorithm grows an ensemble of decision trees without pruning. The idea is to combine many base learners to improve the result, with each learner performing slightly differently due to randomness during the training. The first source of randomness is to train each tree on a randomly selected sub-sample from the original training data. This procedure is called bootstrap aggregating (bagging) and the samples are drawn with replacement. The second possibility to introduce randomness is to randomly select at each node only a subset of predictors. The number of predictors to select for each node was set to the next integer (ceiling function $\lceil\cdot\rceil$) of the square root of the number of features (i.e., $\lceil\sqrt{21}\rceil$). The number of trees was set to 100. The classification decision was based on the unweighted majority vote.

For ML algorithms, the imbalance between OVS and non-OVS in the data set had to be taken into account. To avoid training a classifier that is less sensitive to the minority class, which is in our case the class of interest, hence OVS, we introduced cost-sensitive learning. For the training of the RF, a cost matrix was defined with penalty weights for misclassifications. In the current application, missing OVS (i.e., false negatives) was worse than misclassifying OVS (i.e., false positives). One heuristic approach to defining the misclassification costs is to utilize the imbalance ratio, which is the fraction between the number of instances of the majority class (here non-OVS) and the minority class [36]. However, the current approach considered the estimated class distributions. As described, a large inter-subject variability in OVS (between 7% [11] and 26% [13]) has been reported. Within EMA studies in hearing sciences thus far, mainly elderly participants were included, but the proposed OVD algorithm should be generally applicable. Therefore, an OVS percentage of 12% [10] was assumed as a good compromise. Hence, the false positive cost was set to 0.12 and the default false negative cost of 1 was used. This resulted in an oversampling of the minority class to grow each tree on a more balanced data set. Only a marginal influence of the false positive cost is expected, due to small differences in the real underlying OVS proportion relative to the compromise value of 12%.

Furthermore, since the PSD values were smoothed, the manually labeled data did not fit perfectly in a 125 ms grid pattern. Each positive result was extended with one additional interval before and after the positive classification. A 5-fold cross validation was performed to find the best parameter value for the minimum leaf size, i.e., the minimum number of observations per terminal node. To take the small amount of OVS into account, Matthews correlation coefficient (MCC) was optimized, which is a reliable quality measure for imbalanced datasets [37]. The optimum MCC of 0.740 was reached with a minimum leaf size of 4. This corresponds to a good balance between high true positive rate (TPR) of 98.1% and acceptable false positive rate (FPR) of 21.5% for the decision whether the own voice was present in a 125 ms interval. A lower FPR for real-world recordings covering a whole day is expected, since the percentage of OVS and other speech was much smaller when compared to the training data. False positive detections were mainly due to other speakers that were rather close to the olMEGA-microphones, or to short, loud and coherent environmental sounds. In some cases non-conversational expressions such as laughing or humming were also detected as OVS. Single conversational OVS-words with a short duration (such as "yes" or "no") and less pronounced OVS were mainly responsible for false negative detections.

2.4 Evaluation measurements

To evaluate the OVS classification algorithm, the first-author of this paper recorded audio and acoustical features covering 11.3 h with olMEGA throughout one day. These real-world recordings took place in a completely independent scenario compared to the training data or the intended application of OVD, and they represent a typical day in the life of the author. All communication partners were asked for their consent. Due to the current pandemic situation, the overall contact rate was relatively low. Nevertheless, the recordings reflected normal life, with family meals, piano lessons, and online learning. The data was manually labeled for speech sources.

Figure 1 shows the percentage of manually labeled OVS for different analysis intervals during the recording day. The analysis intervals were overlapped by 50%, and the results are reported for the center of the analysis interval for an easy comparison. Obviously, the variance increases and higher maximum percentages of OVS occur for shorter interval lengths, since the amount of own-voice articulation pauses increases in longer analysis intervals. In 10 s-intervals, even 90% OVS are possible, which is very unlikely for intervals of 5 min. However, the mean voicing times of one day are independent of the interval length (e.g., per min vs per hour). The recordings show active regions (e.g., family meals) and longer intervals with no OVS. These are mostly periods of studying alone. A global perspective with averaged values may hide this information, but the information could be

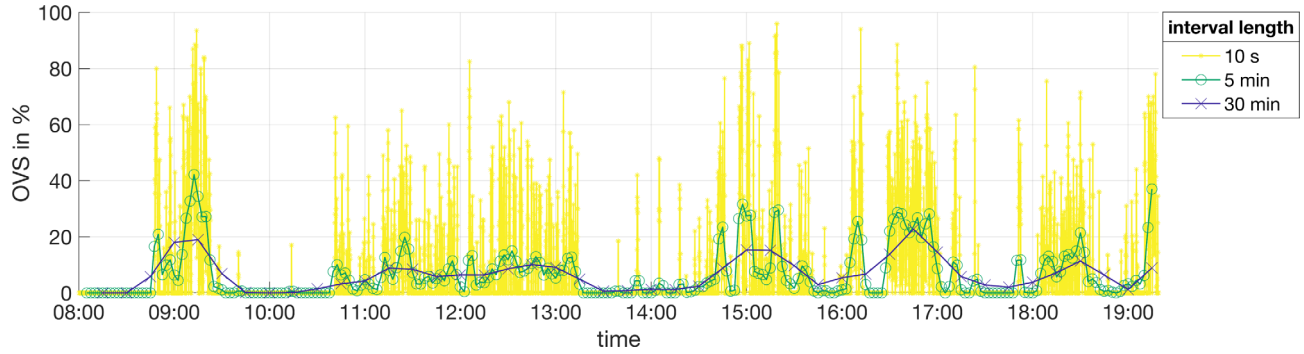

Figure 1. Percentage of manually labeled OVS for different analysis interval lengths during the real-world recordings of one subject.

essential for some applications. Therefore, 10 s-intervals were used for further analysis.

The distribution of the mean SPL for half-overlapping 10 s-intervals L_{10s} is shown in the upper part of Figure 2. On the one hand, only intervals containing OVS were considered, on the other hand all OVS were excluded, and background sounds including other voices were considered. The resulting mean SPL distributions for OVS and background sounds differ clearly in shape. While the background mean SPL were mostly below 60 dB(A), the own voice showed a distribution focus at higher mean SPL. Overall, it can be seen that the recording person spent most of the time in quiet environments. Based on the mean L_{10s} for own voice and background, the local SNR was defined as the ratio between the power of the own voice compared to the power of all other sounds for each interval. If an interval contained no OVS, the local SNR was set to $-\infty$. These intervals were considered neither in the lower part of Figure 2 nor for averaging. For the underlying recording day, the distribution of the local SNR is approx. normally distributed with a mean of 9.4 dB. The local SNR is higher than 0 dB for 95.0% of the intervals.

For this labeled data the proposed OVD using the RF algorithm was applied with one interval extension before and after an interval of 125 ms with predicted OVS.

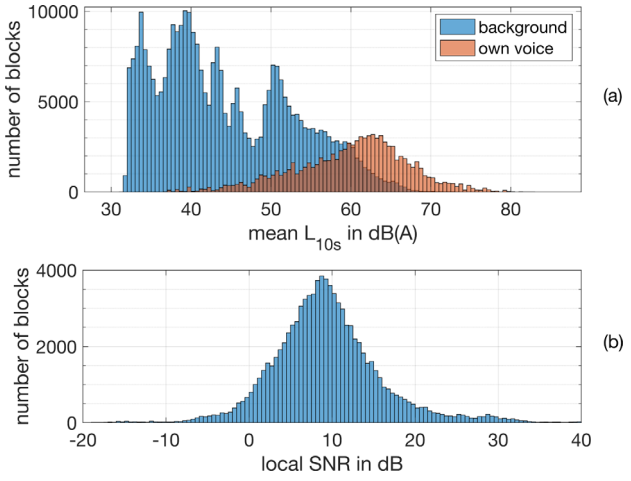

Figure 2. (a) Distribution of mean SPL for half-overlapping 10 s-intervals L_{10s} separated for OVS and all other sounds (background) and (b) the corresponding local SNR during the real-world recordings of one person.

2.5 OVD performance and discussion

For the full-day, real-world recordings, only 7.2% of all manually labeled data were OVS. This value is less than the 12% assumed during training, but it is consistent with the 7% reported for office workers [11] and the 9% for research staff [12]. The proposed OVD algorithm predicted 8.4% OVS, which is rather close to the ground truth percentage. Table 1 shows the confusion matrices for the proposed OVD and the OVD based on the fixed coherence threshold described in [31]. Percent correct for the comparison between manually labeled ground truth and RF-based predictions are given along the diagonal of the confusion matrices (highlighted in bold).

Typical measures as the TPR and FPR (see columns "Est. OVS" in Table 1) were derived to quantify the quality of the classification algorithm. The MCC for the proposed OVD was 0.760, compared to a MCC of 0.625 for the solution based on the fixed coherence threshold [31]. The proposed OVD outperformed the algorithm described in [31] for all measures.

Figure 3 shows situation-specific results for TPR and FPR. The results indicated that the TPR increased with increasing local SNR. Hence, in situations with high local SNR, the proposed OVD was very precise. Those situations contained quiet backgrounds, while own voice is the dominant source. Most of the 10 s-intervals contained no OVS (72.2%). Nevertheless, for these intervals, the FPR was only 1.1%, which is equivalent to 5.53 min incorrectly classified as OVS. In total, 2.5% of all intervals were incorrectly classified as OVS, which is equivalent to 15.93 min. Regarding the total recording duration of 11.3 h, the observed FPR was rated as acceptable and the OVD as robust.

The impact of a false negative and a false positive detection on the mean L_{10s}, compared to the ground truth labels, was analyzed separately. Averaged over the whole day, false negative detections led to a mean difference of -0.3 ± 1.2 dB, and false positive detections to 0.2 ± 1.3 dB, i.e., slightly overestimating and underestimating the mean L_{10s}, respectively. Hence, the impacts of both

Table 1. Detected OVS for real-world recordings of one person. In total, 326 087 intervals with a duration of 125 ms were classified. The manual labels are denoted as "True". OVS detected with the RF algorithm and with the algorithm described in [31] are denoted as "Estimated" (Est.).

	OVD with RF algorithm		OVD described in [31]	
	Est. OVS	Est. non-OVS	Est. OVS	Est. non-OVS
True OVS	**19 660 (84.1%)**	3717 (15.9%)	**16 501 (70.6%)**	6876 (29.4%)
True non-OVS	7645 (2.5%)	**295 065 (97.5%)**	10 738 (3.5%)	**291 972 (96.5%)**

Note: Bold: Percent correct for the comparison between manually labeled ground truth and OVS predictions are given along the diagonal of the confusion matrices.

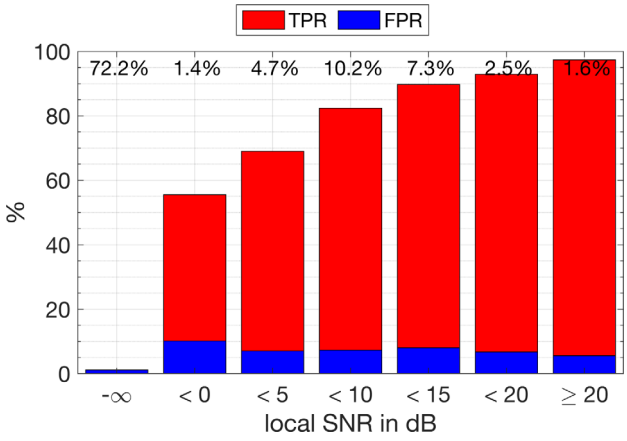

Figure 3. TPR and FPR for the real-world recordings of one person, dependent on the local SNR. A local SNR of $-\infty$ indicates intervals without OVS. The numbers above each bar represent the portion of the recording with the corresponding local SNR.

misclassifications are of opposite sign and the mean difference between ground truth labels and OVD results in 0.04 ± 0.7 dB.

A second 5-fold cross validation was conducted combining the evaluation data and the previously mentioned training data into one big data set to quantify how much the results would variate if other days and/or other persons (e.g., other age, other sex) were considered. Overall, the MCC with the newly trained random forests was on average 0.743, which is between the MCC of 0.740 for the test data in Section 2.3 and the MCC of 0.760 for the evaluation with one day of one person. Hence, there is a small variation in the OVD performance, but the proposed OVD is robust against different speakers and recording environments.

3 Near-ear, long-term study analysis

The second research complex focused on the analysis of real-world NELT-recordings with a larger group of elderly listeners that had hearing impairments. The high variability of collected data required the definition of validity criteria. These criteria were applied, and the predicted proportion of OVS per day and SPL distributions over time were computed.

3.1 Participants and methods

In the analysis, 13 participants (4 females) of the EMA study by von Gablenz et al. [15] were included. All participants were recruited by local hearing aid acousticians prior to the fitting of new hearing aids. All participants had a mild-to-moderate hearing loss. Ten participants wore the olMEGA system without hearing aids. The others were experienced hearing aid users and wore their own hearing aids simultaneously. Their ages ranged from 50 to 75 years, with a median of 65. Eight participants were retired, and three were working full-time.

The participants wore the olMEGA system pre-pandemic for approx. 4 days (range 3–5 days). Only data between 06:00 and 24:00 were considered to exclude times when the olMEGA system most likely was not being worn, but was not switched off. In total, 522.25 h of acoustical data were collected. The mean recording time per participant was 40.17 ± 9.35 h.

The excluded time periods between 24:00 and 06:00 showed mostly low variances in the corresponding PSD data. Similar, longer time periods appeared for some participants during the day, without any subjective assessment. Thus, the combination of stationarity in the PSD and missing subjective assessments is an indication that the olMEGA system was not being worn, e.g., presumably during an afternoon nap when the recordings were erroneously not switched off. By assuming that movement resulted in varying SPL, a minimum-variance criterion for the A-weighted SPL for half-overlapping intervals of 5 min was defined to detect unworn time periods. Based on the comparison of the NELT-data with subjective assessments, the criterion for invalid intervals was empirically set to a standard deviation smaller than 1.5 dB. To have a robust indicator as to whether the system was worn, the minimum-variance decision was smoothed across 13 adjacent half-overlapping 5 min-intervals (i.e., 30 min). The goal was to exclude long unworn periods, at the expense of missing rather short unworn periods. For the considered NELT-data, the unworn decisions seemed plausible. After applying the minimum-variance criterion, the valid total recording time was 493.93 h, or 94.6% of the original recordings.

Additionally, there were situations with a predicted OVS percentage of above 90% in intervals of 5 min, which is very unlikely considering a typical ratio of speech pauses of approx. 17% for non-professional readings and 25% for dialogues [38]. These parts were marked as misclassification

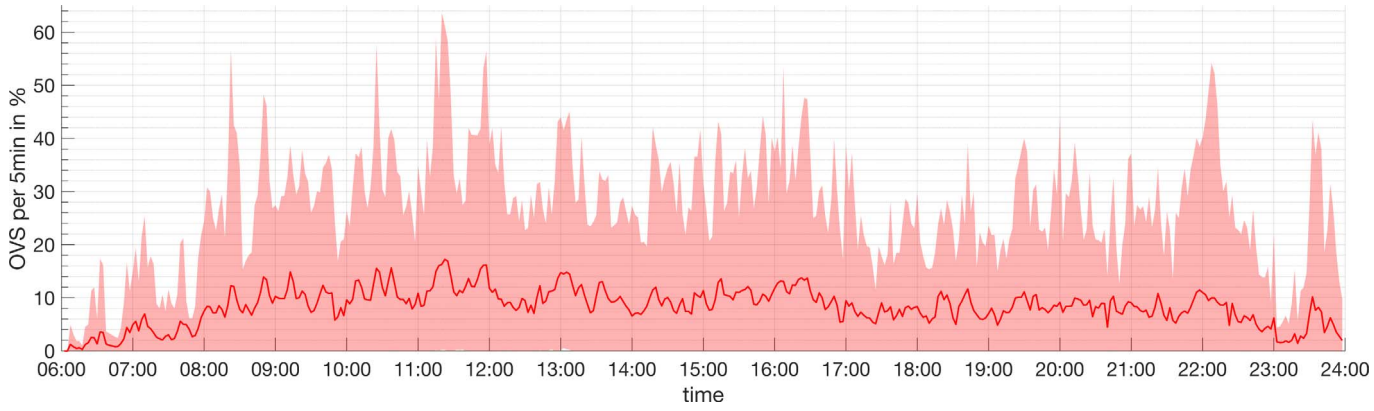

Figure 4. Time course of predicted OVS percentages in intervals of 5 min with 50% overlap, averaged over all participants and recording days. The solid line gives the mean values and the shaded area covers the 10th to 90th percentiles.

and excluded from further analysis, although they might include true positive classifications. The reasons for these invalid intervals were mostly a persistently high coherence between both olMEGA-microphones, e.g., presumably during previously undetected unworn situations when the glasses were folded, or while driving by car and listening to the radio. Only 0.5% of the 5 min-intervals fulfilled this criterion, distributed among different recording days and participants. Their exclusion further reduced the overall analysis time to 491.42 h.

Due to shortcomings in the hardware design of the olMEGA version used, three stationary, sinusoidal disturbances at approx. 5 kHz, 6.5 kHz and 10 kHz had to be removed. The corresponding frequency bins with a bandwidth of approx. 200 Hz were removed for SPL calculations. The removal of these high frequencies had only a marginal effect on the signal of interest, since acoustic SPL are mostly small at high frequencies due to the common $1/f$-characteristic of environmental noise, where f denotes the frequency in Hz. Compared to the overall bandwidth, the applied filter characteristic is very narrow. Subsequently, the hardware design was successfully modified in the current olMEGA version.

3.2 Own voice statistics

For the analysis of the OVS, the questions of interest were whether the participants showed normal conversation behavior in terms of quantity, and whether the amount of OVS was different over the course of the day.

Figure 4 shows the mean percentage of OVS and the 10th and 90th percentiles during the whole day, averaged over all participants and recording days for intervals of 5 min with 50% overlap. The course of the day in Figure 4 shows a slow rising of OVS in the early morning and an increase in activity in the late morning. However, no real rest phases or high-activity phases were common to all participants. Thus, to use OVS in SPL measurements, there was no preferred time period, and no intervals could be excluded.

The grand mean percentage of OVS per day was 10.3%. The individual average percentage of predicted OVS ranged from 5.8% to 17.1%. The mean value of 10.3% was lower than the pre-selected voicing percentage of 12% used in the training process. However, this result is consistent with the voicing percentage of 9.9% reported for elderly retirees [14].

3.3 Level statistics

Figure 5 shows the 10th and 90th percentiles and the median of the A-weighted SPL during the day. Intervals of 5 min with 50% overlap were analyzed and averaged for all participants and recording days. The blue and red curves show the results with or without excluding OVS in the analysis, respectively. The chosen analysis interval of 5 min was quite long, especially if public communication situations are of interest, but appropriate to show a whole day of data, since otherwise the amount of data cannot be displayed appropriately.

The grand median SPL across all participants and recording days was 50.3 dB(A) (s.d. = 12.0 dB). Without A-weighting the grand median SPL was 51.7 dB (s.d. = 12.0 dB), which is close to the grand median SPL of 54.4 dB (s.d. = 6.7 dB) reported by Christensen et al. [9]. The small difference might be explained by differences in the measurement equipment (hearing aids in [9]) or in the amount of noise pollution in a medium-sized German city like Oldenburg and its rural environment. After excluding predicted OVS, the grand median SPL reduced to 49.0 dB (A) (s.d. = 10.9 dB). On average, the differences in median SPL added up to 1.4 dB. However, the 90th percentile was significantly higher if OVS were not excluded from the analysis. This supports the observation that a person's own voice is often the loudest signal close to their ears.

The effect of OVS on the individual equivalent SPL over one day was, for all participants, in the range between 0.1 dB (1.0% OVS) and 6.3 dB (22.6% OVS).

3.3.1 Impact of interval length

However, this global description could be misleading and a more detailed analysis could be of interest for many applications, for example noise-level measurements.

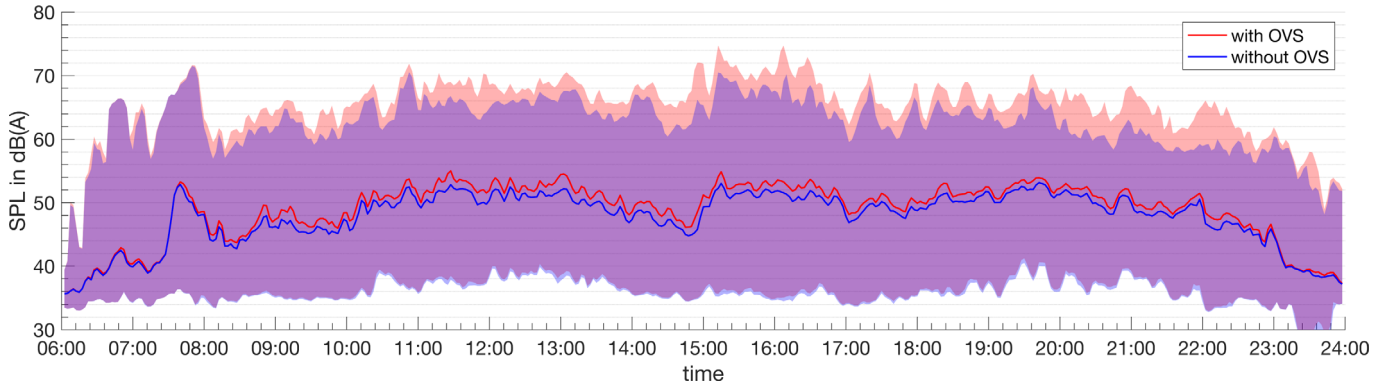

Figure 5. Time course of the A-weighted SPL in intervals of 5 min with 50% overlap, averaged over all participants and recording days. The solid lines give the median values and the shaded areas cover the 10th to 90th percentiles. Results including OVS are given in red and results excluding OVS are given in blue.

Therefore, Figure 6 shows the difference ΔL in the standard SPL measurement (A-weighted mean) due to OVS exclusion. The difference was computed for five interval lengths between 5 s and 300 s for all intervals containing OVS. To improve visibility, the range of the difference in SPL was limited to −10 to 30 dB. Colors indicate the proportion of OVS in each interval. Thus, yellow dots indicate a low proportion and blue dots a high proportion of OVS per 10 s. All intervals with 100% OVS were excluded (i.e., 6425, 1449, 124, and 27 intervals for interval lengths of 5 s, 10 s, 30 s, and 60 s, respectively). The violin plot itself is an approximation of the density distribution for the SPL differences. Hence, a narrow plot shows rare events and a wide one common SPL differences. Furthermore, the median and the 25% and 75% quartiles are shown, as well as the proportion of SPL differences greater than 5 dB.

The violin plots in Figure 6 are wide for small SPL differences, where the proportion of OVS is small. For high proportions of OVS, high SPL differences up to 30 dB and more occurred. The observed range of SPL differences decreased with increasing interval length. However, the results indicate that, depending on the interval length, a relevant proportion of SPL differences is greater than 5 dB. Especially for shorter analysis intervals and a high proportion of OVS, rather high SPL differences can occur. Thus, if background noise or local SNR are the research topic, OVS should be considered.

The observable negative SPL differences seem at first glance to be counter-intuitive, but if the non-stationarity of all signals is taken into account, negative values occur when parts of the background sounds are louder, e.g., by a hand-clap, burst/click-noises, or a communication partner who spoke loudly, compared to the parts with OVS.

3.3.2 Impact of OVS on individual SPL

A different analysis for the data is given in Figure 7. The scatter plot shows the mean A-weighted SPL, including OVS, on the horizontal axis and without OVS on the vertical axis, for all participants and recording days. The mean A-weighted SPL was calculated for 10 s-intervals

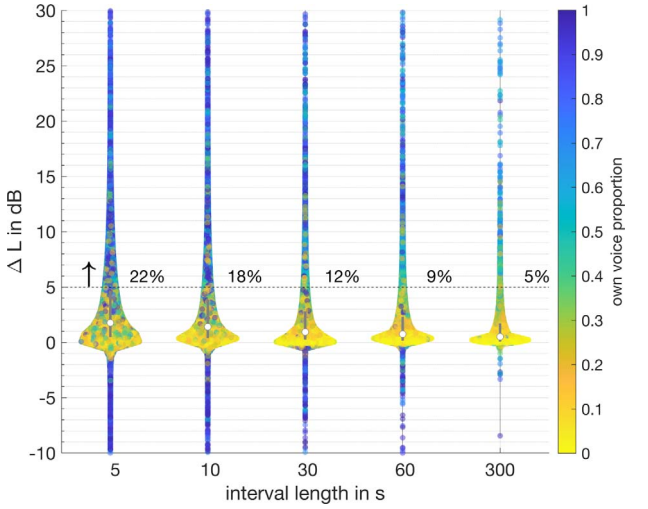

Figure 6. Influence of OVS exclusion on the mean A-weighted SPL in half-overlapping analysis intervals with different interval lengths for all participants and recording days. The color indicates the proportion of OVS in each interval. The median level difference is represented by the white dot and the interquartile range by the grey area. The percentages on the dashed line indicate the proportion of SPL differences greater than 5 dB.

with 50% overlap. The color indicates the proportion of OVS in each interval. Dots on the main diagonal represent no difference in the mean A-weighted SPL due to OVS exclusion.

For most parts of the data, the OVS proportion is rather small and the data are close to the main diagonal. The influence of the OVS is strongest in the mid-level range for SPL of the background between 40 and 60 dB(A) (vertical axis). The own voice increased the measured SPL up to 30 dB. The distance between the corner of the blue cluster and the main diagonal decreases for higher SPL. At SPL above 75 dB(A) the proportion of OVS has only a small impact on the measured SPL. All data points are close to the main diagonal, independent of the proportion of OVS. These results fit well with the observation that we adapt our voice

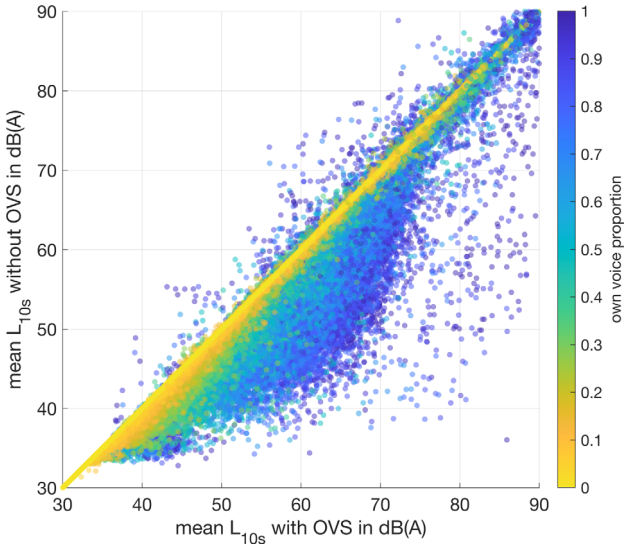

Figure 7. Distribution of the mean A-weighted SPL L_{10s} calculated for 10 s-intervals with 50% overlap for all participants and recording days, on the horizontal axis including OVS and on the vertical axis excluding OVS. The color indicates the proportion of OVS in each interval.

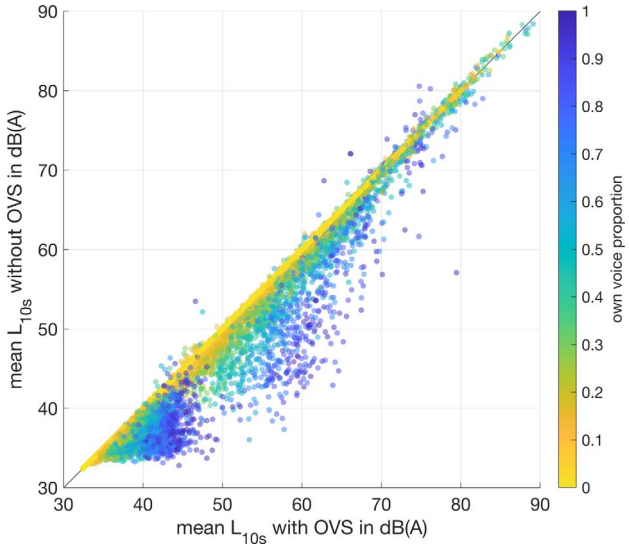

Figure 8. Distribution of the mean A-weighted SPL calculated for 10 s-intervals with 50% overlap for all recording days of one participant, on the horizontal axis including OVS and on the vertical axis excluding OVS. The color indicates the proportion of OVS in each interval. Only this participant showed a cluster of dots with high OVS proportions for rather low SPL.

to the environment, and that this adaptation is reduced for higher noise SPL [39]. The sparse single dots above 75 dB(A), with very high OVS proportions and SPL differences are caused by only one participant. It is assumed that this person had a loud voice but that it is still in the typical range of individual speech SPL [40]. If the data of this participant is excluded from analysis, the grand median SPL with and without OVS differed by less than 0.1 dB.

Below 40 dB(A) background noise, the proportion of OVS and, thus, the distance to the main diagonal decreases. Nevertheless, this assumption cannot explain the left-lower corner of Figure 7. The second dense cluster of dots with high OVS proportions looks like an outlier, since the high proportion of OVS does not fit the overall SPL. This is especially true considering that in quiet environments a typical SPL for male speakers with a normal voice is 54 dB(A), measured at a distance of 1 m to the speaker's mouth. Speaking quietly reduces the SPL to 42 dB(A), and for lower SPL, whispering begins [39]. Typical speech SPL for NELT-recordings were expected to be higher, because of the small distance between the speaker's mouth and the head-worn microphones.

To resolve this deviation, all participants were analyzed separately. Figure 8 shows the analysis of one participant. All data from the questionable lower left corner originated from this participant and only from one day. Further analysis showed that the signal energy was reduced, and a low-pass effect was visible in the PSD. It is assumed that the participant wore the system inside a pocket of his shirt. This hypothesis is supported by the observations that phonation at SPL below 42 dB(A) is rather unlikely, and that on all other recording days of the same person the SPL with OVS was clearly higher. While the recordings of that specific day had no significant influence on the grand median SPL with or without OVS, they could be excluded from all further analysis. However, the scatter plot showed its value for further data analysis and outlier detection.

3.3.3 Impact of background SPL

Ryherd et al. [1] proposed a method for predicting the contribution to a dosimeter measurement of OVS for different amounts of speaking time. To compare our results to Ryherd et al., the mean difference in the A-weighted SPL ΔL due to OVS exclusion was calculated in dependence on the background SPL, i.e., mean L_{10s} without OVS (see Fig. 7) for all participants and recording days. The background SPL was analyzed between 45 dB(A) and 85 dB (A) in intervals of 5 dB, hence each symbol in Figure 9 marks the center SPL. Additionally, the predicted proportion of OVS was divided into five classes from low (5–20%) to high (80–95%). The mean ΔL was highest for low background SPL, combined with a high proportion of OVS, and decreased with increasing background SPL and decreasing proportion of OVS. Our results indicate that, especially at low background SPL, OVS can have a strong influence on the measured SPL. This is consistent with the laboratory measurements by Ryherd et al., although they reported overall larger ΔL. The mean ΔL compared to Ryherd et al. for low background SPL and high OVS percentages (80–95%) is approx. 4 dB lower, and is approx. 8 dB lower for low OVS percentages (5–20%). Hence, our results show a lower impact on the measured SPL, especially for low OVS percentages. Ryherd et al. measured the A-weighted SPL with and without the participant's speech for 1 min for each noise condition, using a dosimeter placed on the participant's shoulder. Their participants

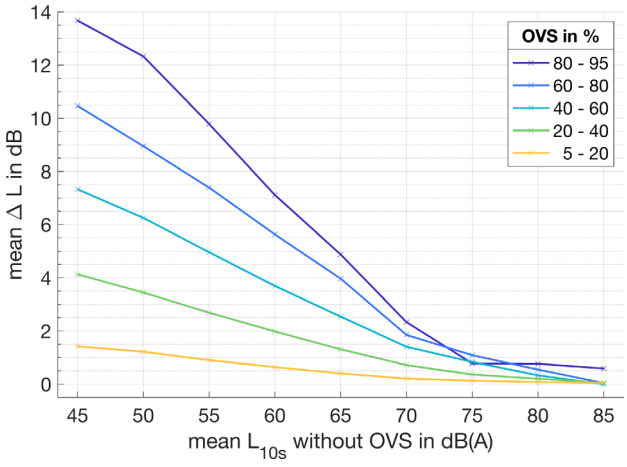

Figure 9. OVS-Exclusion: Mean difference in the A weighted SPL ΔL due to OVS exclusion, calculated for 10 s-intervals with 50% overlap for all participants and recording days. The contours depend on the background SPL (L_{10s} without OVS) and on the predicted proportion of OVS.

read out artificial Hagerman sentences [41], which include rather short Swedish words, to an imaginary communication partner at a distance of 1.6 m. Hence, differences in the speech SPL could occur due to:

- the presence or absence of an actual communication partner (e.g., visual cues),
- the position of the real or imaginary communication partner,
- the comparison of well-defined measurements in the lab to real-world conditions.

4 Conclusions

This contribution showed that long-term observations of SPL considering the participant's own voice are possible without sacrificing the privacy of participants. For near-ear acoustic measurements, the participant's own voice was often the loudest source. Whether it has to be considered depends on the desired application; smaller analysis intervals lead to a high impact. However, for long-term analysis, the OVS impact is rather small and in many cases can be neglected. For the necessary OVD, the proposed random forest classification approach is capable of detecting most of the OVS correctly, and the results compare well with typical speaking times reported in the literature. This is also true for the analysis of the SPL data. Overall, the open-source toolbox and hardware system olMEGA showed its value for EMA. The next step will be a comparison of the objective acoustic data to the assessments of the participants in the corresponding survey app. One additional remaining issue is the development of a SNR estimator based on this system.

Acknowledgments

The authors thank Sascha Bilert, Maximilian Hehl, Florian Schmitt, Nils Schreiber, Kristin Sprenger, and Annäus Wiltfang for data collection, Sven Franz, Holger Groenewold, Sven Kissner, and Ulrik Kowalk for hardware and software development and modification of the operating system, Petra von Gablenz for helpful comments on a draft of this article, and two anonymous reviewers for helpful comments on an earlier version of this article. This work was supported by the Hearing Industry Research Consortium (IRC). English language services were provided by Ute Bitzer and www.stels-ol.de.

Conflict of interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability statement

Data are available on reasonable request from the authors. Part of the data include recordings made by Hendrikse et al. [30] available in Zenodo under the reference https://doi.org/10.5281/zenodo.1621950.

References

1. S. Ryherd, M. Kleiner, K.P. Waye, E.E. Ryherd: Influence of a wearer's voice on noise dosimeter measurements. The Journal of the Acoustical Society of America 131, 2 (2012) 1183–1193.
2. G. Blanken, J. Dittmann, H. Grimm, J.C. Marshall, C.W. Wallesch: Linguistic disorders and pathologies: An international handbook, in Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK), De Gruyter, 2008.
3. I. Holube, P. von Gablenz, J. Bitzer: Ecological momentary assessment in hearing research: Current state, challenges, and future directions. Ear & Hearing 41, S1 (2020) 79S–90S.
4. N.S. Jensen, C. Nielsen: Auditory ecology in a group of experienced hearing-aid users: Can knowledge about hearing-aid users' auditory ecology improve their rehabilitation, in Proceedings of the 21st Danavox Symposium, Kolding, Denmark, August 31–September 2, 2005, 235–258.
5. K.C. Wagener, M. Hansen, C. Ludvigsen: Recording and classification of the acoustic environment of hearing aid users. Journal of the American Academy of Audiology 19, 04 (2008) 348–370.
6. K. Smeds, F. Wolters, M. Rung: Estimation of signal-to-noise ratios in realistic sound scenarios. Journal of the American Academy of Audiology 26, 02 (2015) 183–196.
7. Y.-H. Wu, R.A. Bentler: Do older adults have social lifestyles that place fewer demands on hearing? Journal of the American Academy of Audiology 23, 09 (2012) 697–711.
8. Y.-H. Wu, E. Stangl, O. Chipara, S.S. Hasan, A. Welhaven, J. Oleson: Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss. Ear & Hearing 39, 2 (2018) 293–304.
9. J.H. Christensen, G.H. Saunders, M. Porsbo, N.H. Pontoppidan: The everyday acoustic environment and its association with human heart rate: Evidence from real-world data logging with hearing aids and wearables. Royal Society Open Science 8 (2021) 1–16.

10. I.R. Titze, E.J. Hunter, J.G. Svec: Voicing and silence periods in daily and weekly vocalizations of teachers. The Journal of the Acoustical Society of America 121, 1 (2007) 469–478.
11. T. Masuda, Y. Ikeda, H. Manako, S. Komiyama: Analysis of vocal abuse: Fluctuations in phonation time and intensity in 4 groups of speakers. Acta Oto-Laryngologica 113, 4 (1993) 547–552.
12. D. Mehta, H. Cheyne, A. Wehner, J. Heaton, R. Hillman: Accuracy of self-reported estimates of daily voice use in adults with normal and disordered voices. American Journal of Speech-Language Pathology 25 (2016) 1–8.
13. M.R. Mehl, J.W. Pennebaker, D.M. Crow, J. Dabbs, J.H. Price: The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations. Behavior Research Methods, Instruments, & Computers 33, 4 (2001) 517–523.
14. N. Fhärm, F. Skoglund, J. van Doorn: Time spent talking in retirement, in Proceedings of the 15th Australasian International Speech Science and Technology Conference, Christchurch, New Zealand, 2014.
15. P. von Gablenz, U. Kowalk, J. Bitzer, M. Meis, I. Holube: Individual hearing aid benefit in real life evaluated using ecological momentary assessment. Trends in Hearing 25 (2021) 1–18.
16. A. Carullo, A. Vallan, A. Astolfi: Design issues for a portable vocal analyzer. IEEE Transactions on Instrumentation and Measurement 62, 5 (2013) 1084–1093.
17. T. Powers, M. Froehlich, E. Branda, J. Weber: Clinical study shows significant benefit of own voice processing. Hearing Review 25, 2 (2018) 30–34.
18. T. Behrens, C. Nielsen, T. Lunner, C. Elberling: Method of programming a communication device and a programmable communication device. US Patent 7,340,231, 2008.
19. M. Lugger: Hearing apparatus with own speaker activity detection and method for operating a hearing apparatus. US Patent 8,873,779, 2014.
20. V. Hamacher: Hearing apparatus and a method for own-voice detection. US Patent 7,853,031, 2010.
21. S. Granqvist: The self-to-other ratio applied as a phonation detector for voice accumulation. Logopedics Phonatrics Vocology 28, 2 (2003) 71–80.
22. K.B. Rasmussen, S. Laugesen: Method for detection of own voice activity in a communication device. US Patent 7,512,245, 2009.
23. J. Bitzer, S. Bilert, I. Holube: Evaluation of binaural own voice detection (OVD) algorithms, in Speech Communication; 13th ITG-Symposium, VDE, 2018, 1–5
24. J. Bitzer, S. Kissner, I. Holube: Privacy-aware acoustic assessments of everyday life. Journal of the Audio Engineering Society 64, 6 (2016) 395–404.
25. S. Kissner, I. Holube, J. Bitzer: A smartphone-based, privacy-aware recording system for the assessment of everyday listening situations, in Proceedings of the International Symposium on Auditory and Audiological Research, Nyborg, Denmark, August 26–28, 2015, 445–452.
26. U. Kowalk, S. Kissner, P. von Gablenz, I. Holube, J. Bitzer: An improved privacy-aware system for objective and subjective ecological momentary assessment, in Proceedings of the International Symposium on Auditory and Audiological Research, Nyborg, Denmark, August 23–25, 2017, 25B–30B.
27. R. Baker, V. Hazan: DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. Behavior Research Methods 43, 3 (2011) 761–770.
28. G. Grimm, J. Luberadzka, V. Hohmann: A toolbox for rendering virtual acoustic environments in the context of audiology. Acta Acustica United with Acustica 105, 3 (2019) 566–578.
29. M.M.E. Hendrikse, G. Llorach, V. Hohmann, G. Grimm; Movement and gaze behavior in virtual audiovisual listening environments resembling everyday life. Trends in Hearing 23 (2019) 1–29.
30. M.M.E. Hendrikse, G. Llorach, V. Hohmann, G. Grimm: Virtual Audiovisual Everyday-Life Environments for Hearing Aid Research (Version 2) [Database]. Zenodo, 2019. https://doi.org/10.5281/zenodo.1621950.
31. J. Bitzer, S. Kissner: Two-channel coherence-based own voice detection for privacy-aware longterm acoustic measurements, in Speech Communication; 12. ITG Symposium; Proceedings of, VDE, 2016, 1–5.
32. S. Graf, T. Herbig, M. Buck, G. Schmidt: Features for voice activity detection: A comparative analysis. EURASIP Journal on Advances in Signal Processing 2015, 1 (2015) 91–105.
33. S. Schötz: Acoustic analysis of adult speaker age, in M. Christian, Ed. Speaker Classification I, Lecture Notes in Computer Science, Vol. 1, Springer, 2007, 88–107
34. L. Breiman: Random forests. Machine Learning 45, 1 (2001) 5–32.
35. G. Biau, E. Scornet: A random forest guided tour. TEST 25, 2 (2016) 197–227.
36. A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera: Cost-sensitive learning, Springer International Publishing, Cham. 2018, 63–78.
37. S. Marsland: Machine learning: An algorithmic perspective. CRC Press, 2009.
38. S. Gustafson-Čapková, B. Megyesi: A comparative study of pauses in dialogues and read speech, in Proceedings of Eurospeech 2001, Aalborg, Denmark, 2001, 931–935
39. H. Lazarus, C.A. Sust, R. Steckel, M. Kulka, P. Kurtz: Akustische Grundlagen sprachlicher Kommunikation. Springer, 2007.
40. K.S. Pearsons, R.L. Bennett, S.A. Fidell: Speech levels in various noise environments, Office of Health and Ecological Effects, Office of Research and Development (EPA-600/1-77-025) US Environmental Protection Agency, Washington DC, 1977.
41. B. Hagerman: Sentences for testing speech intelligibility in noise. Scandinavian Audiology 11 (1982) 79–87.

Cite this article as: Pohlhausen J. Holube I. & Bitzer J. 2022. Near-ear sound pressure level distribution in everyday life considering the user's own voice and privacy. Acta Acustica, **6**, 40.