OPEN ACCESS Freely available online

Journal of the Audio Engineering Society
Vol. 64, No. 6, June 2016 (© 2016)
DOI: http://dx.doi.org/10.17743/jaes.2016.0020

Privacy-Aware Acoustic Assessments of Everyday Life*

JOERG BITZER, SVEN KISSNER , AND INGA HOLUBE
(joerg.bitzer@jade-hs.de) (sven.kissner@jade-hs.de) (inga.holube@jade-hs.de)

Jade University for Applied Sciences, Institute of Hearing Technology and Audiology, D-26121 Oldenburg, Germany

There is not much data available on the acoustical properties of our everyday environments and our subjective impressions and communication abilities in those environments. In this paper we present a smartphone-based system that is capable of measuring the most important features of the surrounding acoustics. However, in contrast to simple audio recordings and subsequent analysis the proposed solution will respect the privacy of all communication partners and bystanders by extracting a feature set tested for privacy compliance only. Nonetheless, a larger set of necessary features for acoustical environment classification can be computed in the necessary accuracy afterwards in an off-line process. For a given feature set the statistical analysis shows comparable results in the extracted data, when either the original audio signals or the new privacy-aware extraction methods are used.

0 INTRODUCTION

People are surrounded by acoustical environments everywhere and at all times. How they behave in those environments can be subject to factors such as mood or capabilities. In order to enhance people's possibilities to interact with their environment, hearing devices are common tools. However, it is difficult to evaluate the benefit of those tools or to measure the actual acoustically challenging situations in natural environments. In the research focus "Hearing in everyday life in Oldenburg (HALLO)" one important task is to assess the typical acoustic environment of everyday situations. Since the subjects are not supervised and they interact with people not involved in the study, special care has to be taken to protect the privacy of everyone involved, including bystanders and their conversations.

In order to monitor people's behavior and their perception, many studies are based on questionnaires that are filled in afterwards in the test laboratory or completed in the evening at home (e.g., [15, 3]). Unfortunately, these methods are prone to subjective or even wrong memories. In order to reduce the time between a given situation and the questionnaire, the method of ecological momentary assessment (EMA) was introduced. EMA allows to get ongoing data through self-reports, often implemented as the experience sampling method (ESM) [8], e.g., for hearing aid users [4]. The results are still on subjective scales.

A natural choice to do acoustic assessments is to sample the environment acoustically. Combined with the data obtained by EMA, this enables comparisons of objective measures to the respective subjective ratings. Several methods are known from very different research fields. In [18] and [11] binaural audio recordings were used to measure noise exposure, phonation time [5], and vocal effort. Conversation behavior was analyzed in [13] by recording only small fragments of audio (30 s every 12.5 min) and allowing participants to exclude recordings from further analysis, though only a very few used those options [12]. The combination of questionnaires and objective data, especially on mobile phones, was proposed in [17] e.g., but without considering audio. A technically simpler approach was to carry out interviews and recordings of natural acoustics at the same time [16]. EMA with long-term audio recordings on a mobile phone was proposed by [7, 6].

Yet, none of these approaches will protect the privacy of conversation partners or bystanders by design. The respective national laws on how to deal with this fact are quite different and we can only give some information on US and German law. In the US the recording of conversation is defined by federal law (18 U.S.C. §2511) [2] and further defined by the different states' laws. Most of the US states (e.g., in Texas or Georgia where [13] was conducted) and the federal law allow recording if at least one party gives consent, even if it is the person using the recording device (*One Party Consent*). In some others like California,

*Supported by governmental funding initiative "Niedersächsisches Vorab" of the Lower Saxony Ministry for Science and Culture, research focus "Hören im Alltag Oldenburg (HALLO)"

all participants of a conversation must agree to the recording (*Two Party Consent*). In Germany a two-party consent is necessary (StGB §201) [1]. Another problem arises in public or semi-public spaces, like restaurants or trains. The device could record conversations of bystanders who are not actively a part of the conversation of the participant. In this case no informed consent is given. However, the spoken word in public is protected differently. In Germany the intention of the conversation is important. If it was meant to be private, recording would be forbidden. This shows that for a system which is suited to be used worldwide, very rigid privacy standards have to be implemented.

In this paper we propose a solution based on the extraction of a limited set of features on the smartphone that does not compromise privacy, and yet still allows for the assessment of the acoustical environment. The paper is organized as follows. In the next section the system, its hard- and software, is described and the technical specification is given. Secs. 2 and 3 show the privacy-aware feature extraction method and its evaluation. The final section includes an analysis of the system, to show that the privacy of others is not affected.

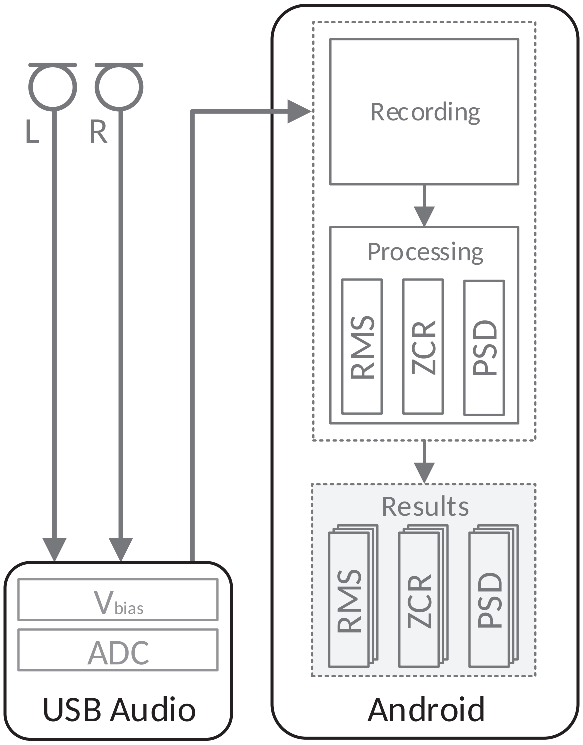

Fig. 1. Schematic of the Android recording system. Shown are the two microphones connected to the USB-Audio-Interface that supplies a bias voltage of 2.2 V (V_{bias}) to power the microphones. The A/D converted audio signal is passed to the smartphone where features are extracted and stored.

1 SYSTEM DESCRIPTION

The full system consists of multiple hardware components as well as control software and analysis algorithms implemented in both the mobile recording device itself and in MATLAB for further off-line processing.

1.1 Hard- and Software

While a number of mobile multichannel recording devices are available, their lack of flexible audio processing capabilities makes it necessary to save the audio material until it can be processed off-line, disqualifying them for continuous long-time recording, if we want to protect the subject's personal data and privacy. Modern smartphones, on the other hand, are affordable and offer sufficient processing power for digital audio processing, allowing to extract selected features and then discard the audio data itself.

Given its openness and flexibility, Android was the platform of choice. Since we were unable to acquire an Android device that itself provides stereo audio input, external USB audio interfaces of type PureAudio USB-MA (Andrea Electronics) were used. Android 5 offers some basic access to external USB audio interfaces but is limited in functionality and device support. For a more flexible solution, a class-compliant driver (i.e., does not require vendor-specific drivers) is commercially available and was purchased to reduce development time. Generally, any USB class compliant device should be compatible. The Android device in question has to support USB On-The-Go, which enables it to act as a host to external USB devices. For this study smartphones of type Moto G (1st generation, 2013, Motorola) were used. Two omni-directional microphones of type EK-23024 (Knowles Electronics) were built into simple behind-the-ear shells and the necessary bias voltage of 2.2 V was supplied by the USB audio interface. Fig. 1 shows an overview of the system's main hard- and software components. Of course different microphones could be used as well, e.g., consumer devices aimed specifically at recording for stereophony. However, those often come in in-ear variants, obstructing the auditory canal and are therefore not suited for the task at hand when a subject should be able to act as unhindered as possible in everyday life.

The software is comprised of a simple user interface (UI) for starting and stopping the analysis as well as a status indicator (Fig. 2). In the background, a service controls signal recording and processing. Once started, the service runs independently from the UI. This allows for ongoing recording and processing, should the user exit the UI itself or change to another application.

Two channels of data, left and right, are sampled simultaneously and recorded continuously with a sample rate of $f_s = 16$ kHz at a resolution of 16 bits and cached in chunks of 60 s. After a chunk is written, the service is notified to start a new processing thread. There, the cached data is high-pass filtered ($f_0 = 100$ Hz, 2nd-order Butterworth) to reduce low-frequency noise (see Fig. 4), passed on to the algorithms as specified in Sec. 3.1, and the resulting feature-data is stored. After the processing of a given chunk of data is finished, the system deletes the cached audio data permanently. With the algorithms currently deployed, the smartphone's battery lasts for over 8 hours, producing 130 MB of feature-data per hour.

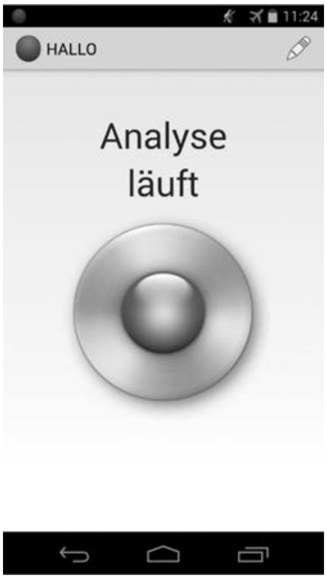

Fig. 2. User interface of the analysis app, indicating that the analysis is running (in German: *Analyse läuft*). Analysis is started and stopped by touching the large button in the middle of the screen.

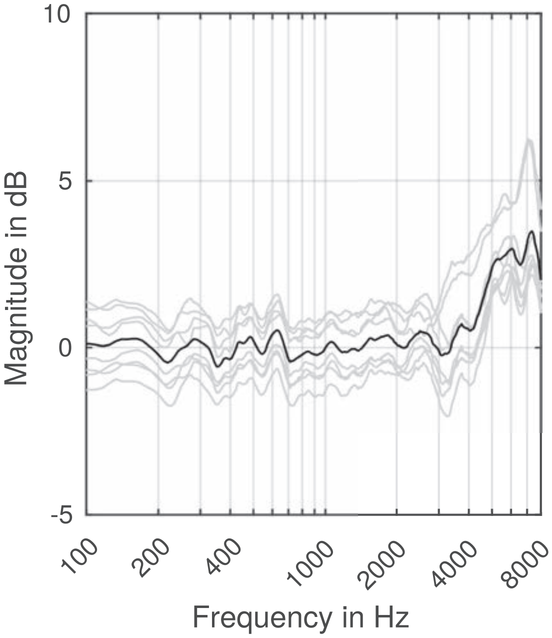

Fig. 3. Frequency response of four measurement systems (8 mics.). Shown is the average response over all microphones (black) as well as the individual transfer functions (grey). The data has been smoothed in ERB-bands for display.

1.2 Measurements

In order to test if the recording system can be used for EMA from a purely technical point of view, the frequency response and the noise floor of four pairs of microphones in their respective behind-the-ear shells were measured. Finally, the dynamic range was determined for each pair.

The measurements to determine the system's characteristics were conducted in the anechoic chamber at our department. Free field responses were measured using a G.R.A.S 40AF free-field microphone with a type 26TK preamplifier powered by Bruel & Kjær type 2829. The signal sound pressure level was calibrated using a Bruel & Kjær calibrator, type 4231 (114 dB SPL at 1 kHz).

1.2.1 Frequency Response

To evaluate the frequency response of the microphones, Gaussian white noise with a broadband level of 74 dB SPL was played back using an NTi Audio Talk-Box. This loudspeaker, developed as a reference speaker for room acoustic applications, offers on-line equalization of any input signal resulting in a flat frequency response between 100 Hz and 10 kHz with a steep roll-off beyond. It was mounted at a distance of 1.75 m to the investigated microphone.

The power spectral densities (PSD) were estimated from the recorded white noise using Welch's method ([19]) and smoothed using a moving average filter with equivalent rectangular bandwidth (ERB, [14]). These PSDs were then related to an equally processed PSD derived from the reference microphone, yielding the transfer functions as shown in Fig. 3. The black line denotes the average over all microphones whereas the grey lines show the individual transfer functions.

Up to 3 kHz, the responses are flat within 1 dB. Between microphones, the differences are within 2 dB, showing very similar progression. Beyond 3 kHz, there is a slight rise, and while six microphones remain within 2 dB of each other, two microphones show a moderately increased sensitivity.

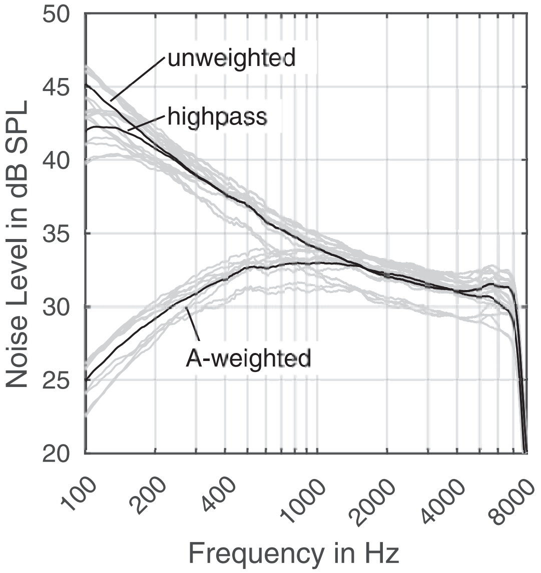

Fig. 4. Noise level as a function of frequency for unweighted, highpass-filtered and A-weighted noise floor. Shown is the average over all microphones (black) as well as the individual measurements (grey). The results have been smoothed in ERB-bands for display.

1.2.2 Noise Floor

The noise floor was determined by relating measurements of the static background noise to calibrated white noise measurements from above. Again, the results

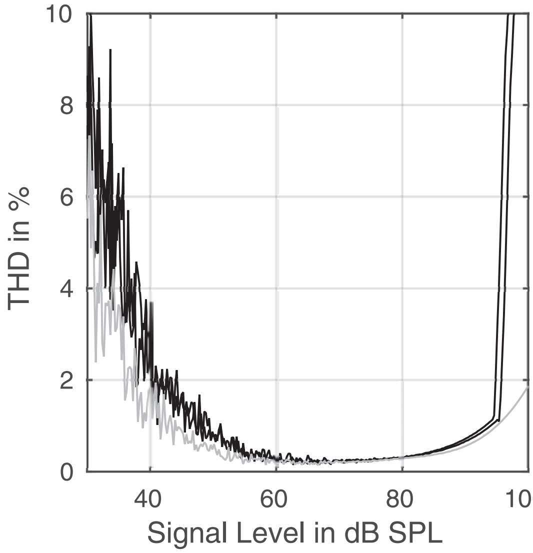

Fig. 5. Total harmonic distortion as a function of sound pressure level for a 1 kHz sine. Depicted are two channels of one measurement system (black) and the same signal recorded with a reference microphone (40AF, grey).

have been smoothed in ERB-bands. In addition to the unweighted noise-level, a second order high pass filter corresponding to the implementation on the Android system ($f_0 = 100$ Hz, 2nd-order Butterworth) was applied as well as an A-weighting filter. The resulting noise levels are shown in Fig. 4.

As with the frequency response, the microphones exhibit very similar noise floors. One pair, the same that showed a slightly higher sensitivity towards higher frequencies above, shows a noise floor about 2 dB below the other microphones. Generally, this noise floor below 35 dB SPL beyond 1 kHz hints at a viable dynamic range, since all relevant acoustic situations we would like to acquire are far above 35 dB SPL. A normal conversation, for example, can be assumed to take place at levels of at least 50 dB SPL.

1.2.3 THD and Dynamic Range

The dynamic range of the system is given by the noise floor at the low end and by the clipping behavior at the upper end of the range. In order to determine the upper bound of the system, we measured the total harmonic distortions (THD) using an amplitude swept sine with a fixed frequency of 1 kHz. To produce appropriate sound levels of up to 100 dB SPL, a Fostex 6301B active loudspeaker was used in close proximity to the microphones (0.6 m). The THD was calculated by relating the amplitude of the second to fifth harmonics (V_2 through V_5, without the fundamental frequency) to the amplitude of the first harmonic at 1 kHz, V_1,

$$\mathrm{THD} = \frac{\sqrt{V_2^2 + V_3^2 + V_4^2 + V_5^2}}{V_1}. \tag{1}$$

Fig. 5 shows the THD as a function of the signal level for one pair of microphones (black) as well as for the reference microphone (grey). For the analysis only levels above the static noise level of ≈34 dB SPL (compare Fig. 4) are considered. The THD slowly decreases as the signal level rises above the noise floor, showing an increasingly smoother progression. At around 95 dB SPL, the recording system abruptly reaches its limit for transparent audio capture (clipping) while the reference microphone shows a moderately increasing THD. Considering a THD of below 1 or even 2% as sufficiently low for acoustic transducers, this results in a usable dynamic range of 45 to 55 dB. The results of the other three microphone pairs and measurement systems are similar.

2 FEATURE EXTRACTION

The final application for the analysis of everyday life recordings will define the subset of features and parameters that are necessary to be extracted from the audio signal. The proposed system is very flexible and can compute many different features. However, due to power constraints on a smartphone and privacy considerations not everything that is desirable can be done.

We will focus on a feature subset proposed by [9] for acoustic classification in hearing aids. It includes, e.g., descriptors for the power of the signal, the zero-crossing rate, or the Mel Frequency Cepstral Coefficients (MFCC).

In order to apply the extraction guidelines beneficially, we first analyzed which parts have to be computed on the audio stream and which parts can be derived afterwards from other extracted data or features. Thus, the computational and storage demands are reduced on the smartphone system. The final dataset is computed afterwards off-line in a PC-based framework. Additionally, for all extraction routines, privacy is considered.

Many features are based on some form of frequency domain description. We decided to extract the power spectral densities (PSD) for the left and right signal and the cross-power spectral density (CPSD) to get information on the spatial acoustics, which extends Kates' original approach.

Therefore, our final solution is not an exact replica of the extraction routines given by Kates but a close approximation in order to measure and classify acoustic environments.

2.1 Smartphone Extraction Process

We refer to the recorded audio channels as $x_\mathrm{L}[n]$ and $x_\mathrm{R}[n]$, respectively. To streamline the expressions below, formulae are given for one channel $x[n]$ unless noted differently.

Before processing, a 2nd-order high-pass Butterworth filter with a cutoff-frequency of $f_0 = 100$ Hz is applied to the signal in order to suppress low-frequency noise.

The processing itself is frame-based, with each block x divided into M overlapping frames x_m with $N = 400$ samples each (25 ms) and a feed of $L = N/2 = 200$ samples (12.5 ms).

$$x_m[n] = x[n + m \cdot L], \tag{2}$$

with current frame number $m = 0, 1, \ldots, M - 1$ and samples $n = 0, 1, \ldots, N - 1$.

For each frame the Root Mean Squared (RMS) value and the Zero Crossing Rate (ZCR) is computed. Furthermore, the ZCR of the first derivation of the input signal $\Delta x_m[n] = x_m[n] - x_m[n-1]$ is determined.

The power spectral densities (PSD) are not in the Kates-set, but they are valuable to derive other features and they are important for our privacy-aware system. Hence, we will explain the extraction process in more detail.

The PSDs are estimated by first applying a Hann-window

$$\omega[n] = 0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] \tag{3}$$

to signal frames $x_{\mathrm{L},\,m}$ and $x_{\mathrm{R},\,m}$. Afterwards, the windowed frame is transformed to the frequency domain using a 512-point fast Fourier transform (FFT). Based on the frequency-domain representations $X_{\mathrm{L},\,m}$ and $X_{\mathrm{R},\,m}$, the periodogram for the left and right channel, Φ_{L}^{Per} and Φ_{R}^{Per} as well as the Cross-periodogram Φ_{LR}^{Per} are computed:

$$\Phi_{[\mathrm{L,R}],\mathrm{m}}^{Per}[n] = X_{[\mathrm{L,R}],m}[n] \cdot X_{[\mathrm{L,R}],m}^{*}[n], \tag{4}$$

$$\Phi_{\mathrm{LR}}^{Per}[n] = X_{\mathrm{L},m}[n] \cdot X_{\mathrm{R},m}^{*}[n], \tag{5}$$

where $(\cdot)^*$ denotes the complex conjugation. A PSD estimate is finally achieved by smoothing the adjacent periodograms. For non-stationary processes like speech usually a first order recursive filter

$$\hat{\Phi}[m] = \alpha \cdot \Phi^{Per}[m] + (1-\alpha) \cdot \hat{\Phi}[m-1], \tag{6}$$

is applied, where

$$\alpha = \exp(-t/\tau), \tag{7}$$

denotes the smoothing factor, with frame-shift $t = (N - L)/f_s$ (12.5 ms) and time constant τ. Since speech signals contained in the recorded audio data $x_{\mathrm{L}}[n]$ and $x_{\mathrm{R}}[n]$ can easily be reconstructed to the level of intelligibility from the original periodograms, we have to apply a time constant as high as 125 ms. Additionally, every 125 ms the current averaged frame with all three spectra is saved, the rest of the data is discarded to ensure privacy (see listening test, Sec. 4).

2.2 Feature Extraction in the Off-Line Post Processing

The saved feature data are used to compute several other features given by Kates. All of them are very briefly described and notable differences in the extraction procedure compared to the description in [9] are given.

The signal envelope is based on the extracted RMS values and therefore no differences between the direct audio and the extracted features are to be expected, since both are working on the same frame interval of 12.5 ms.

As a second feature the well-known Mel Frequency Cepstral Coefficients (MFCC) are derived from the PSD estimates. They are often used for speech recognition tasks and are well suited for classification purposes. They represent the spectral shape of the audio signal in a very compact form.

Other spectral features include the power spectral centroid and the power spectrum entropy. In this special implementation the results are given in the critical band numbers (used for the MFCCs) and not in Hz.

The extraction of the broadband correlation is somewhat more complex. It can be divided into the audio-dependent part, which is the sum of powers of the critical bands without the low and high frequencies. The result per audio block is smoothed to get an envelope, and for this envelope a sliding autocorrelation function is computed. The maximum and the lag-value are the two final features. These features can be used to find repetitive structures in the envelope. They cannot be compared directly to the audio-extracted material and the smartphone-based features, since the time interval varies too much (12.5 ms compared to 125 ms). Thus, we additionally extracted the broadband version of these features by using the RMS values and comparing these results with the audio-based features.

In contrast to these signal-based extraction methods, the delta coefficient is used directly on the extracted feature. It can be computed for all other features and is defined as the difference between two adjacent time instances. It represents the rate of change of the feature structure. Another statistical measure derived is the standard deviation, estimated by subtracting a running mean that is computed with a first order low-pass filter and a time constant of 500 ms. Finally, the running standard deviation is computed.

3 MEASUREMENTS

For the ZCR and the PSD estimates we will show some results of test routines. All derived features will be compared to the same features computed directly from the input audio signal.

3.1 Zero Crossing Rate

The zero crossing rate (ZCR) was evaluated from pure tones at 1/3-octave center frequencies, played back in an anechoic chamber using the TalkBox (see Sec. 2.2), and simultaneously recorded. Fig. 6 shows boxplots of the combined data of both channels, equalling 6 s of analysis per frequency or 478 frames for a frame-size of 25 ms with 50% overlap. Generally there is little deviation of the median from the theoretical values with a maximum of 4% at 125 and 250 Hz. The analysis at 8 kHz shows an increased number of outliers, which can be explained by distortions or artifacts caused by the proximity to the Nyquist-frequency.

3.2 Power Spectral Densities

From the same sine sequence as used above we estimated $\hat{\Phi}_{\mathrm{L}}$, $\hat{\Phi}_{\mathrm{R}}$ and $\hat{\Phi}_{\mathrm{LR}}$ according to Eq. (3)ff. and calculated the absolute power per frequency relative to full scale. The results, as functions of time, are shown as spectrograms in Fig. 7. The sine sequence can easily be traced with each tone at the expected frequency. The faint harmonics are also visible. At lower frequencies the noise contained in the signal increases. This is in accordance with the system's design and the noise measurements shown in Fig. 4.

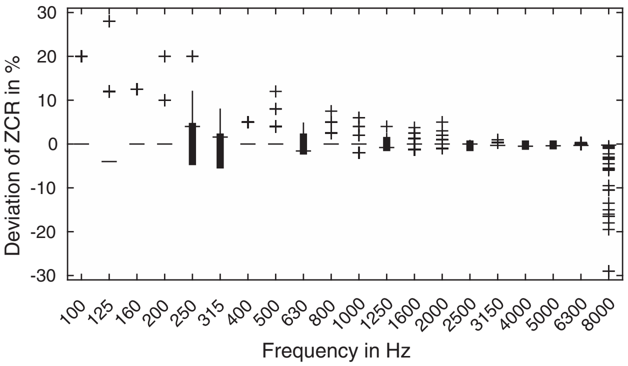

Fig. 6. Deviation of the zero crossing rate as calculated by the Android system from the theoretical value for each frequency displayed. The box plots show the median (horizontal line), the lower and upper quartile (lower and upper boundary of the box), the lowest and highest values within 1.5 times the quartile range relative to the lower and upper quartile (limit of vertical line), and outliers (symbol +).

Both measurements confirm that the basic features are extracted according to the design.

3.3 Measurement and Comparison of the Off-Line Features

In order to show a comparison of the direct audio recoding to the privacy-aware extraction we recorded three different acoustic scenarios: a silent office, the same office with people talking in it, and a busy street with lots of cars passing by.

Fig. 8 shows the four MFCCs for the three situations (10 s each) extracted by using the original audio signal (black) and by using the proposed privacy-aware method (grey).

The results clearly show that the plots for the new method are much smoother and therefore some details are lost. However, the overall trajectories are quite similar and each situation is clearly different from the others. The statistical analysis of 40 s of audio material of each situation confirms these results (see Figs. 9 to 12). The boxplots reveal slight differences in the statistics, which is to be expected since the number of data points is different by a factor of 10 (12.5 ms and 125 ms block interval). However, the three situations can be statistically separated for both analysis methods.

Other features, especially the delta measures, differ more between the two methods. An example is given[1] in Fig. 13. Due to the different block intervals the quantile range diverges, yet all of the three scenarios result in discriminable statistics.

4 PRIVACY ISSUES AND LISTENING TEST

The smartphone system extracts RMS and ZCR values and PSDs. Only the PSDs can be used to reconstruct speech in such detail that the resulting signal is intelligible. Therefore, we decided to use smoothed PSDs only. This section describes the listening experiment to determine the recognition score of speech reconstructed from the stored PSDs in dependence of the smoothing constant τ (25, 75, and 125 ms, see Eq. (7)).

4.1 Test Design

Speech recognition was determined by using the Göttingen sentence test (GÖSA, [10]). This test consists of 200 everyday sentences combined in 10 test lists with 20 sentences each spoken by a male speaker. The sentences are composed of three to seven words per sentence. The speech material was used in its original version as well as in the three processed versions. One test list was presented to each listener for each version over headphones (HDA200) using the software program Oldenburg Measurement Application (OMA) of HörTech GmbH, a Fireface 400 (RME), headphone driver HB7 (Tucker Davis Technologies), and a MicroAmp HA400 (4-channel Stereo Headphone Amplifier, Behringer). The OMA was modified to a "user-defined speech test" to include the processed speech material. The speech material was presented without background noise (in quiet) at a level of 70 dB SPL. One list of the original Göttingen sentence test was presented and the listener used the volume control of the MicroAmp HA400 to adjust the level of the presented signals. After the listener found the level for highest perceived recognition, the presentation of the list was interrupted. Then, a second list with original speech was presented and the listener repeated the presented sentences. The examiner marked each correctly recognized word on a touch screen. The measurements were continued with the processed versions in a randomized order. The list was also randomly selected.

Ten young normal hearing listeners (seven male, three female, age 20–27 years) participated in the experiment. Their hearing loss was 10 dB HL at maximum from 125 Hz to 4 kHz and 20 dB HL at maximum until 8 kHz.

4.2 Signal Generation

The test signals were generated by calculating averaged PSDs from the original GÖSA sentences according to Eq. (3)ff, with a frame size of 25 ms and an output rate of 25 ms (every frame), 75 ms (every third frame), and 125 ms (every fifth frame). The frames were smoothed by using the corresponding smoothing factor τ (see Eq. (7)).

Afterwards we calculated the minimum phase φ for each frame from its Hilbert-transform,

$$g[m] = -\mathcal{H}\{\log(|\hat{\Phi}[m]|)\}, \tag{8}$$

$$\varphi[m] = \mathrm{Im}\{g[m]\}, \tag{9}$$

where $\log(\cdot)$ denotes the natural logarithm and $\mathrm{Im}\{\cdot\}$ the operator to retrieve the imaginary-part only. $\varphi[m]$ is then applied to $\hat{\Phi}[m]$ and the spectrum is transformed to the time domain. To prevent transients, a Hann-window is applied to each frame. Finally, we adjusted the level to get the output signal. The resulting speech is reconstructed to the same length as the original, with continuous data-frames for the

[1] In this paper only a minor selection of the many features are presented. You can download all figures, the extraction code, and the original data at the corresponding web-page (http://pub.tgm.io/PrivacyAwareEMA_2016).

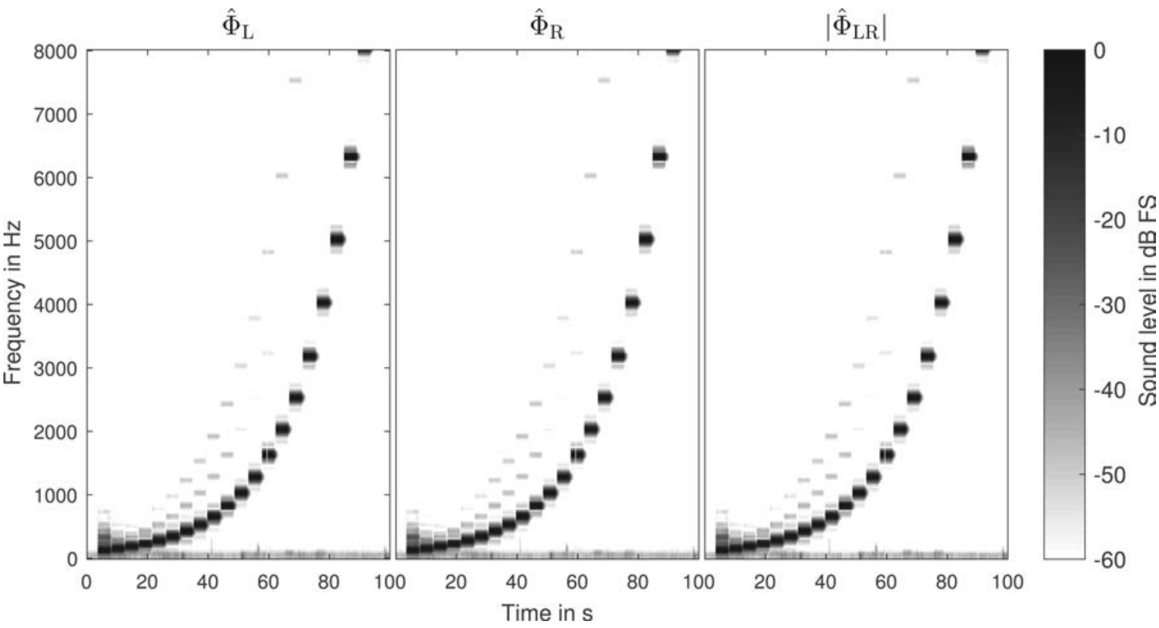

Fig. 7. Spectrograms calculated from the power spectral densities for the left and right channel ($\hat{\Phi}_L$, $\hat{\Phi}_R$) as well as from the magnitude of the cross-power spectral density ($|\hat{\Phi}_{LR}|$), estimated for a sequence of sine tones at 1/3-octave center frequencies from 100 Hz to 8 kHz.

output rate of 25 ms, whereas for 75 and 125 ms the signal is padded with two and four 25 ms frames of zeros between data-frames respectively.

4.3 Results and Discussion

Fig. 14 shows the recognition scores in percent correct for words in the Göttingen sentence test. The original sentences are fully recognizable resulting in a score of 100% for all listeners. The processed sentences with a window length of 25 ms result in a recognition score of 97.2% (median). Two listeners still reach a score of 100%. With broadening the window length, the recognition score drops to median values of 1.6% for 75 ms and 0.6% for 125 ms. Five out of the 10 listeners were not able to repeat one single word of the respective test list in the 125 ms version. The highest score for this version was 3.2%. Therefore, this parameter setting can be regarded as "privacy-proof." This goal is also almost achieved by a window length of 75 ms. Nevertheless, the Göttingen sentence test offers a speech rate of 279 syllables per minute and is therefore regarded as a relatively fast speech test. The recognition rate might be higher for slower speaking communication partners in everyday life. Hence, the processing with a window length of 125 ms was used for further testing.

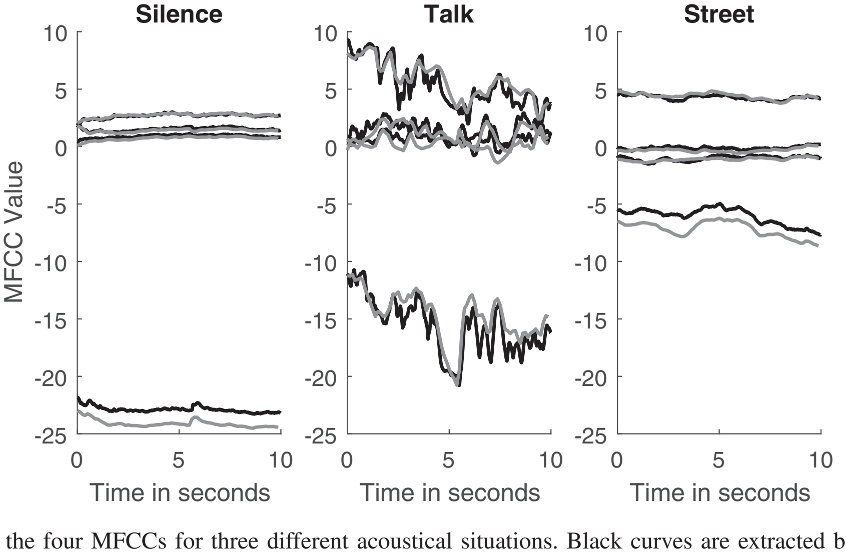

Fig. 8. Comparison of the four MFCCs for three different acoustical situations. Black curves are extracted by using the audio data directly. Grey lines are derived by the proposed privacy-aware method. Note that we added an arbitrarily chosen constant of 30 to the lowermost line to reduce the dynamic of the whole figure.

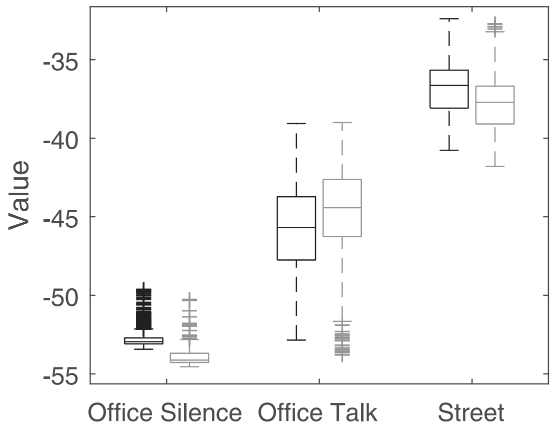

Fig. 9. Boxplot of the first MFCC for three different acoustical situations. Black plots (left of each pair) are extracted by analyzing the audio data directly. Grey plots are derived by the proposed privacy-aware method.

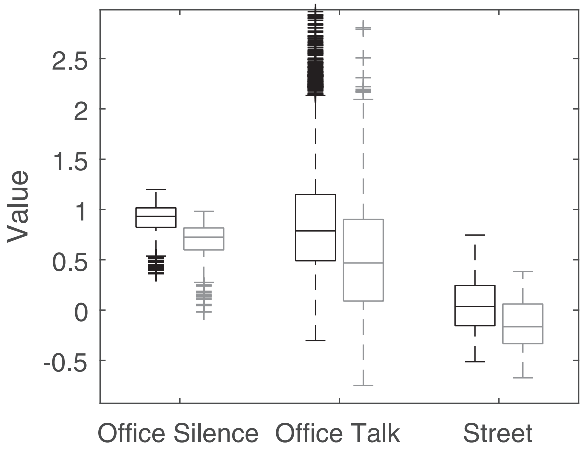

Fig. 12. Boxplot of the fourth MFCC for three different acoustical situations. Black plots (left of each pair) are extracted by analyzing the audio data directly. Grey plots are derived by the proposed privacy-aware method.

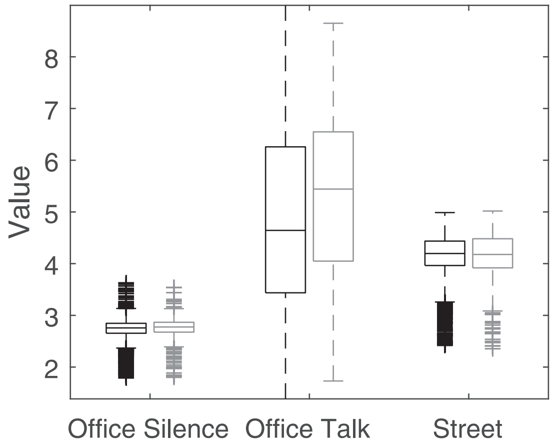

Fig. 10. Boxplot of the second MFCC for three different acoustical situations. Black plots (left of each pair) are extracted by analyzing the audio data directly. Grey plots are derived by the proposed privacy-aware method.

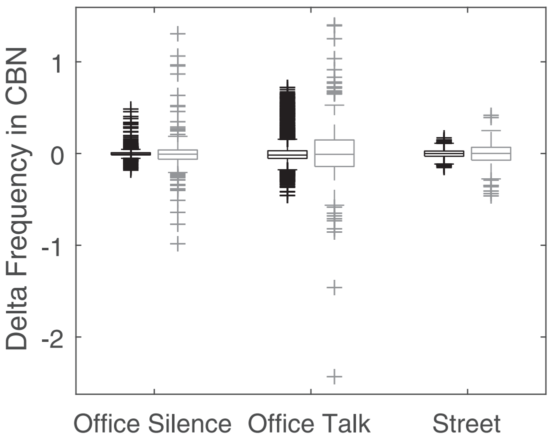

Fig. 13. Boxplot of the delta centroid for three different acoustical situations. Black plots (left of each pair) are extracted by analyzing the audio data directly. Grey plots are derived by the proposed privacy-aware method.

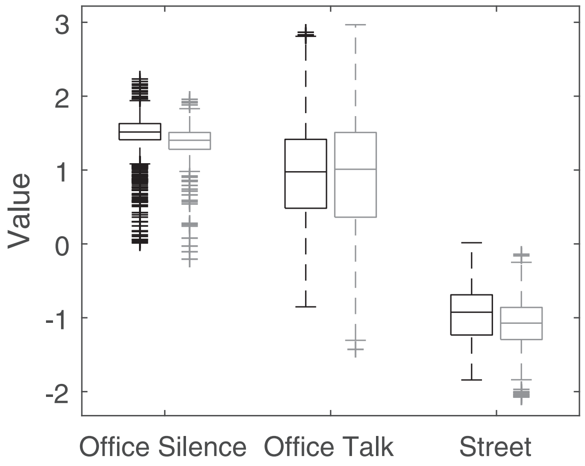

Fig. 11. Boxplot of the third MFCC for three different acoustical situations. Black plots (left of each pair) are extracted by analyzing the audio data directly. Grey plots derived by the proposed privacy-aware method.

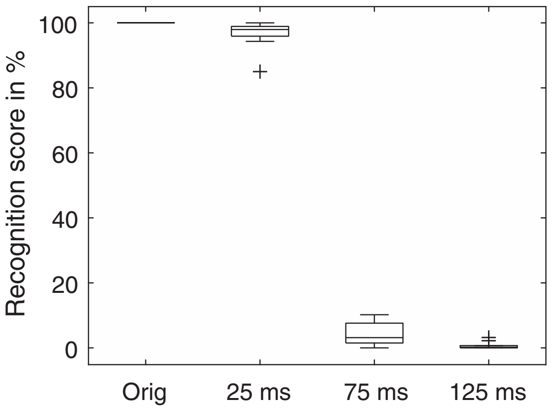

Fig. 14. Recognition score for the original Göttingen sentences and the processed sentences with window lengths of 25, 75, and 125 ms.

5 CONCLUSIONS AND OUTLOOK

In this paper a novel solution to assess the acoustic environment while preserving people's privacy was proposed. We showed that the computational power needed on a smartphone device could be reduced by dividing the feature extraction into two parts: a first extraction phase on a smartphone and a second phase on a more powerful computer platform afterwards. The comparison shows that different scenarios result in separable features for the new extraction method. Therefore, we believe that acoustic assessment without disturbing the privacy of the speaker, conversation partner, or bystanders is possible. The next steps in our research will be to use the system to record material of a few test persons over a period of several days to get better insight into acoustically challenging situations for hearing aid users. Furthermore, new features to describe acoustical situations will be developed, such as estimators for the signal-to-noise ratio and the reverberation time.

The Android program, the source code of the extraction methods (java classes and matlab), and all data are available on the corresponding web page.

6 ACKNOWLEDGMENT

Technical support and data collection of Anne Schlueter and Christoph Eike is gratefully acknowledged. Thanks to the company audifon for providing the hearing aid housings with the ear-worn microphones.

7 REFERENCES

[1] Strafgesetzbuch (StGB) §201 Verletzung der Vertraulichkeit des Wortes. http://dejure.org/gesetze/StGB/201.html.

[2] Electronic Communications Privacy Act of 1986 (ECPA): 18 U.S.C. §2511, Interception and disclosure of wire, oral, or electronic communications prohibited (1986). http://codes.lp.findlaw.com/uscode/18/I/119/2511.

[3] R. M. Cox and G. C. Alexander "The Abbreviated Profile of Hearing Aid Benefit," *Ear and Hearing*, vol. 16, no. 2, pp. 176–186 (1995). http://dx.doi.org/10.1097/00003446-199504000-00005.

[4] G. Galvez, M. B. Turbin, E. J. Thielman, J. A. Istvan, J. A. Andrews, and J. A. Henry "Feasibility of Ecological Momentary Assessment of Hearing Difficulties Encountered by Hearing Aid Users," *Ear and Hearing*, vol. 33, no. 4, pp. 497–507 (2012). http://dx.doi.org/10.1097/AUD.0b013e3182498c41.

[5] S. Granqvist, "The Self-to-Other Ratio Applied as a Phonation Detector for Voice Accumulation," *Logopedics Phonatrics Vocology*, vol. 28, no. 2, pp. 71–80 (2003). http://dx.doi.org/10.1080/14015430310011772.

[6] S. S. Hasan, O. Chipara, Y.-H. Wu, and N. Aksan "Evaluating Auditory Contexts and Their Impacts on Hearing Aid Outcomes with Mobile Phones," in *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, pp. 126–133, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2014). http://dx.doi.org/10.4108/icst.pervasivehealth.2014.254952.

[7] S. S. Hasan, F. Lai, O. Chipara, and Y.-H. Wu "Audiosense: Enabling Real-Time Evaluation of Hearing Aid Technology In-Situ," in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pp. 167–172, IEEE (2013). http://dx.doi.org/10.1109/cbms.2013.6627783.

[8] J. M. Hektner, J. A. Schmidt, and M.y Csiksgentmihalyi *Experience Sampling Method: Measuring the Quality of Everyday Life* (Sage, 2007).

[9] J. M. Kates *Digital Hearing Aids* (Plural Pub., 2008).

[10] B. Kollmeier and M. Wesselkamp "Development and Evaluation of a German Sentence Test for Objective and Subjective Speech Intelligibility Assessment," *J. Acous. Soc. Amer.*, vol. 102, no. 4, pp. 2412–2421 (1997). http://dx.doi.org/10.1121/1.419624.

[11] F. Lindstrom, K. P. Waye, M. Södersten, A. McAllister, and S. Ternström "Observations of the Relationship between Noise Exposure and Preschool Teacher Voice Usage in Day-Care Center Environments," *J. Voice*, vol. 25, no. 2), pp. 166–172 (2011). http://dx.doi.org/10.1016/j.jvoice.2009.09.009.

[12] M. R. Mehl and S. E. Holleran "An Empirical Analysis of the Obtrusiveness of and Participants' Compliance with the Electronically Activated Recorder (Ear)," *European J. Psychological Assessment*, vol. 23, no. 4, pp. 248–257 (2007). http://dx.doi.org/10.1027/1015-5759.23.4.248.

[13] M. R. Mehl, J. W. Pennebaker, D. M. Crow, J. Dabbs, and J. H. Price "The Electronically Activated Recorder (Ear): A Device for Sampling Naturalistic Daily Activities and Conversations," *Behavior Research Methods, Instruments, & Computers*, vol. 33, no. 4, pp. 517–523 (2001). http://dx.doi.org/10.3758/bf03195410.

[14] B. C. J. Moore and B. R. Glasberg "Suggested Formulae for Calculating Auditory-Filter Bandwidths and Excitation Patterns," *J. Acous. Soc. Amer.*, vol. 74, no. 3, pp. 750–753 (1983). http://dx.doi.org/10.1121/1.389861.

[15] W. Noble, N. S. Jensen, G. Naylor, N. Bhullar, and M. A. Aneroid "A Short Form of the Speech, Spatial and Qualities of Hearing Scale Suitable for Clinical Use: The SSQ12," *Intl. J. Audiology*, vol. 52, no. 6, pp. 409–412 (2013). http://dx.doi.org/10.3109/14992027.2013.781278.

[16] M. Raimbault, C. Lavandier, and M. Bérengier "Ambient Sound Assessment of Urban Environments: Field Studies in Two French Cities," *Applied Acoustics*, vol. 64, no. 12, pp. 1241–1256 (2003). http://dx.doi.org/10.1016/s0003-682x(03)00061-6.

[17] N. Ramanathan, F. Alquaddoomi, H. Falaki, D. George, C.-K. Hsieh, J. Jenkins, C. Ketcham, B. Longstaff, J. Ooms, J. Selsky, et al., "Ohmage: An Open Mobile System for Activity and Experience Sampling," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on*, pp. 203–204, IEEE (2012). http://dx.doi.org/10.4108/icst.pervasivehealth.2012.248705.

[18] A. Szabo, B. Hammarberg, S. Granqvist, and M. Södersten "Methods to Study Pre-School Teachers' Voice at Work: Simultaneous Recordings with a Voice Accumulator and a DAT Recorder," *Logopedics Phoniatrics Vocology*, vol. 28, no. 1, pp. 29–39 (2003). http://dx.doi.org/10.1080/14015430310010863.

[19] P. D. Welch "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging over Short, Modified Periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73 (1967). http://dx.doi.org/10.1109/TAU.1967.1161901.

THE AUTHORS

Joerg Bitzer

Sven Kissner

Inga Holube

Joerg Bitzer received his diploma in 1995 and his doctorate in electrical engineering in 2002 from the University of Bremen where he also worked as a research assistant until 1999. From 2000 to 2003 he was head of the algorithm development team at Houpert Digital Audio, a company specialized in audio signal processing. Since September 2003 he is a professor for audio signal processing at the Jade University of Applied Science Wilhelmshaven/Oldenburg/Elsfleth. In 2010 he joined the Fraunhofer project group for hearing, speech, and audio technology in Oldenburg as a scientific supervisor. His current research interests include all forms of single- and multichannel speech enhancement, audio restoration, audio effects for musical applications, and information retrieval for large media archives.

•

Sven Kissner received his B.Eng and M.Sc in hearing technology and audiology from Jade University of Applied Sciences and the Carl von Ossietzky University in Oldenburg respectively. He is currently back at the Jade University as a research associate to work on various projects involving (virtual) acoustics, signal processing, and mobile devices.

•

Inga Holube started studying physics at Göttingen University, Germany, in 1984 and obtained her postgraduate degree (Ph.D.) from the Third Physical Institute at Göttingen University in 1993, which was awarded by the department of physics. Prof. Dr. B. Kollmeier supervised her work and between 1993 and 1995 she worked as a research associate in his team at the University of Oldenburg. During a one-year postdoctoral scholarship from the German Research Council (DFG) she visited research institutions in the Netherlands, the US, and the UK. From 1995 to 2001 she was head of the Audiological Research and System Technology department at Siemens Audiologische Technik GmbH in Erlangen, Germany. Inga Holube received her approval for Medical Physics and the certificate for Audiology of the German Society for Medical Physics. Since September 2001 she is professor for the Hearing Technology and Audiology study course at Oldenburg/Ostfriesland/Wilhelmshaven University of Applied Sciences now reorganized to Jade University of Applied Sciences. Her research interests are subjective and objective measures for the assessment of speech intelligibility and listening effort and the evaluation of hearing instruments as well as the epidemiology of hearing impairment. In 2014 Inga Holube was awarded with a 3-year research professorship from the state of Lower Saxony, Germany.