A smartphone-based, privacy-aware recording system for the assessment of everyday listening situations

SVEN KISSNER*, INGA HOLUBE, AND JOERG BITZER

Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany

When trying to quantify hearing difficulties in every-day listening situations, mostly questionnaires are used to collect and evaluate subjective impressions. Obtaining objective data outside a laboratory is relatively difficult, given the required equipment and its proper handling as well as privacy concerns emerging from long-term audio recordings in a non-regulated and populated environment. Therefore, a smartphone-based system was developed that allows long-term ecological momentary assessment. Microphones are placed close to the ears to obtain signal characteristics, e.g., interaural level differences, similar to those perceived by a listener. Currently, root-mean-square, averaged spectra and the zero crossing rate are calculated. Additional features can be implemented and the flexibility of the smartphone itself allows for additional functionality, e.g., subjective ratings on predefined scales. A simple user interface ensures that the system can be properly handled by non-tech-savvy users. As only the extracted features but not the audio-data itself are stored, screening and approval of the recorded data by the test subject is not necessary. Furthermore, additional standard features, e.g., the spectral centroid, can be computed offline, utilizing the recorded features.

INTRODUCTION

Capturing an acoustical environment regarding its physical characteristics as well as how it is perceived by a subject can be a valuable tool. It allows for improving existing hearing systems, the refinement of fitting procedures and their evaluation, as well as studies on individual experiences and behavior in given situations. In practice, however, situations are often assessed using questionnaires retroactively or under laboratory conditions. Without objective measurements, it is difficult to establish a proper relation between a situation and its perception. Delayed feedback can lead to biased and vague results and a controlled environment does not necessarily reflect real-life conditions or evoke similar reactions. To circumvent those issues, data has to be captured directly in the respective situations (Ecological Momentary Assessment; Shiffman *et al.*, 2008). If a study does aim to capture objective data in-situ, the handling of more or less user-friendly technical equipment can frustrate subjects. To ensure privacy, screening of recorded data and/or a declaration of consent from all parties involved is required, the former being time consuming while the latter often is impractical in public spaces.

*Corresponding author: sven.kissner@jade-hs.de

Proceedings of ISAAR 2015: Individual Hearing Loss – Characterization, Modelling, Compensation Strategies. 5th symposium on Auditory and Audiological Research. August 2015, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C. Dalsgaard, L. Tranebjærg, and T. Andersen. ISBN: 978-87-990013-5-4. The Danavox Jubilee Foundation, 2015.

We developed a system that overcomes the problems described above to a certain extent. The primary goal was a mobile recording system that is easy to use, even for non-tech-savvy users. It should be relatively inexpensive to build multiple devices and make them easily replaceable if needed, not relying on specific components. Programming should be flexible and functionality easily extendable if required. Last but not least, the system is engineered with a subject's privacy in mind.

HOW IT WORKS

The main goal is a system which is able to retrieve certain acoustic parameters from everyday listening situations without storing the audio data itself. Following is a description of the hard- and software (also see Fig. 1) as well as an overview of the chosen features.

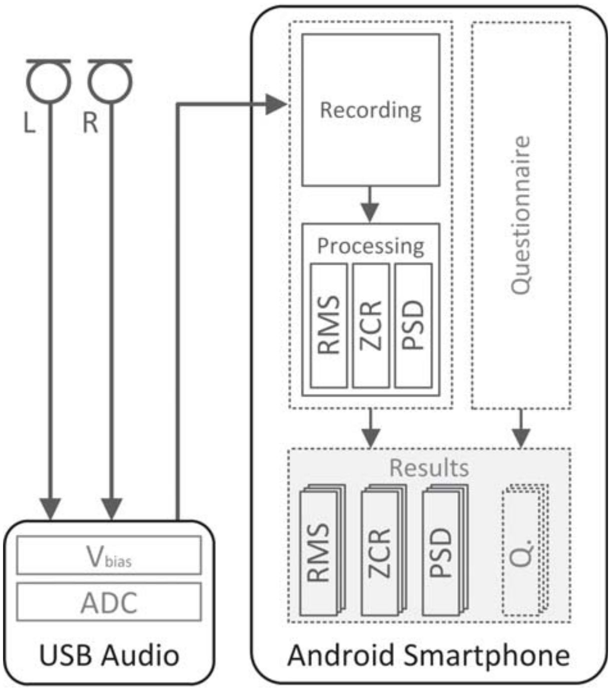

Fig. 1: Schematic of the recording system outlining its hardware components as well as the basic operational sequence.

Hardware

The system consists of three main hardware components. An Android-based smartphone, a USB audio interface, and hearing aid dummies. The Android platform was chosen due to its openness and flexibility. There is a multitude of smartphones and tablets available, fitting almost every conceivable requirement. Unfortunately, most Android-based consumer hand-held devices do not support stereo-input using the standard 3.5-mm audio jack. To circumvent this, we decided to use an external

USB audio interface. This limits the number of eligible devices as the Android-device in question has to support USB-OTG, i.e., can act as a host to USB devices. On the other end, the USB audio device must be class compliant and therefore not require proprietary drivers. Additionally, the interface has to support stereo input, while many audio interfaces with the required small form factor only offer one input channel.

We selected the Moto G by Motorola, an affordable, mid-range smartphone as well as thumb-sized USB audio interfaces of type USB-MA by Andrea Communications.[1] The interface was fitted with an micro-USB port to allow for a direct connection to the smartphone without the need for an additional adapter. The behind-the-ears hearing aid dummies each house a microphone of type EK-23024 by Knowles Electronics. The audio interface supplies a bias voltage of 2.2 V, so no additional power source is required. The dummies are connected to the audio interface using a 3.5-mm stereo audio jack.

Software

The software used for data acquisition and processing was developed in Java using the Android SDK. Currently there is no sufficiently sensible and robust way to access external USB audio interfaces using the API defined by Google. To circumvent the time-consuming implementation of a driver, a third-party product based on libusb was purchased (Dr-Jordan-Design).

Upon opening the Android app, the user is presented with a simple and clean user interface featuring a large button to start or stop data analysis. The button as well as text indicate the current status. The core of the app is a background service. The user interface connects to an existing service or starts a new instance if none is running. Due to the way Android handles lifecycles of an app or activity, a service enables processes to run continuously, even if another app (i.e., camera or questionnaire) becomes active. The service manages audio acquisition as well as processing which both run parallel in their respective threads. Raw audio data is recorded continuously, with a sampling rate of 16 kHz and cached in chunks of 60 s. Each completed chunk is reported to the service which in turn starts a new thread processing the cached data, i.e., sequentially calculating the implemented features and writing the data to the device's storage. Each processed chuck is again reported to the service which deletes the cached audio chunk and, if available, starts processing the next. This is repeated until analysis is stopped and the cached audio data is processed and discarded.

Due to the influence of low-frequency noise the signal is high-pass filtered ($f_0 = 100$ Hz, 2nd-order Butterworth) before processing.

Features of interest

For the selection of acoustical parameters we focused on Kates (2008), who discusses various features for acoustic classification in hearing aids. We also considered *computational complexity* of a certain feature. The system should not back up on cached data, i.e., processing of a single chunk should be finished before the next is completely cached. This also affects power consumption and therefore the maximum duration of continuous recording sessions before the device has to be recharged. As the calculated features are stored on the device, the available *storage space* must be taken into acount as well. A test subject should be able to use the system autonomously for a given amount of time, i.e., four days, eight hours each day, without the need need to daily retrieve the data to free up space. The phone used in the current system provides about 5.5 GB of usable flash-storage. The features currently implemented generate about 130 MB of data per hour, allowing for roughly 40 hours of data. This, of course, can be mitigated by current smartphones offering more internal memory or support for external memory. Another aspect in feature selection is the ability to *derive additional features offline* to save both processing power and storage space.

Therefore, three features being calculated on the recording device. The broadband power (root-mean-square; RMS), the zero crossing rate of the signal and its derivative (ZCR / ΔZCR; Kates, 2008), as well as spectral information in the form of power spectral density (PSD) of left and right channels as well as their cross power spectral density (CPSD; Welch, 1967). Table 1 shows the parameters used for feature extraction.

	RMS	ZCR/ΔZCR	PSD/CPSD
Blocksize in ms	25	25	25/125*
Overlap in ms	12.5	12.5	12.5
nFFT in samples	●	●	512

Table 1: Processing parameters for the selected features. (* the smoothed block is equivalent to 125 ms, see Sec. “Privacy”)

PERFORMANCE

To determine the performance of the system, we measured transfer function, noise level, and total harmonic distortions in an anechoic chamber. To eliminate the speaker’s transfer function (NTi TalkBox and Fostex 6301B) as well as to obtain reference measurements in silence, a G.R.A.S 40AF free-field microphone was used. The results are shown in Fig. 2. Noise floor and frequency response are smoothed in equivalent-rectangular-bandwidth for display.

The frequency responses are reasonably flat within 1 dB up to 3 kHz, rising slightly beyond. The noise of the various sensors also behaves similarly, up to 45 dB SPL at low frequencies (unfiltered), falling below 35 dB SPL beyond 1 kHz. Additionally, the noise is shown after high-pass filtering ($f_0 = 100$ Hz, 2nd order Butterworth) as applied before feature extraction, as well as with A-weighting.

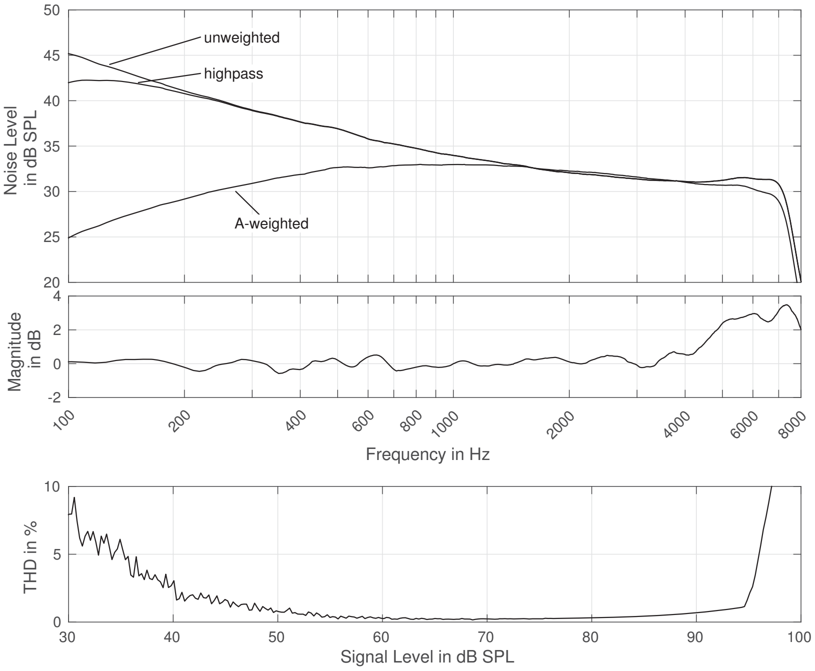

Fig. 2: *Top:* Transfer function. *Middle:* Noise level as a function of frequency for the unweighted, highpass filtered, and A-weighted noise floor, mean over all microphones. *Bottom:* Total harmonic distortion as a function of sound pressure level for a 1-kHz sine. Mean over all measurement systems (10 microhones)

The total harmonic distortions, measured using a 1-kHz amplitude swept sine and calculated from the first four harmonics relative to the fundamental frequency, are dominated by the noise for low frequencies. Beyond a signal level of 40 dB SPL, the signal emerges from the noise until clipping starts abruptly at around 95 dB SPL. The 40AF shows a similar behavior for low levels while it exhibits an extended dynamic with a shallow increase of THDs towards 100 dB SPL. Considering noise levels and THD, the system offers a usable dynamic range of 45 to 55 dB.

Figure 3 shows the selected features, RMS, ZCR and ΔZCR, as well as PSD, as calculated by the system for different situations. Depicted is only one channel. Seconds 0 to 20 show the results for an office with normal background noise, running computers, typing, etc. Seconds 20 to 40 show the results for the same office with two people conversing. Seconds 40 to 60 show the results walking besides a road with steady traffic. The PSD shows increased levels for low frequencies in the quiet office, corresponding with low-frequency microphone noise. Specific events are clearly visible and correspond well over all features, be it keystrokes and speech in the office or loud vehicles passing and a steady background noise for the traffic situation.

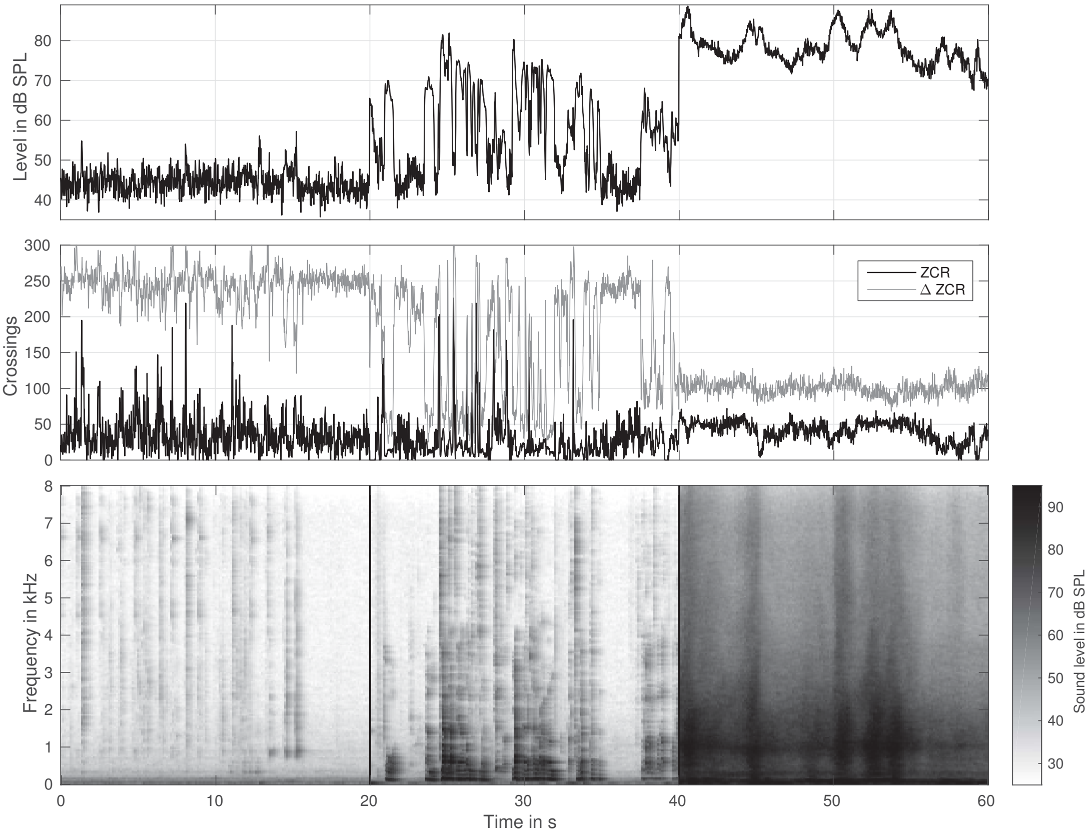

Fig. 3: Examplary RMS (top), ZCR and ΔZCR (middle), as well as PSD (bottom) for one channel as calculated from different situations. *Seconds 0-20:* The author's office with normal background noise, typing, etc. *Seconds 20-40:* A conversation between the author and a collegaue in the same office. *Seconds 40-60:* Walking besides a road with steady traffic.

PRIVACY

While we do not store audio data with respect to a subject's and third-parties' privacy, taking a closer look at the extracted features shows that while broadband RMS and zero crossing rate contain no privacy sensitive information, PSDs are more revealing. With little effort, we are able to reconstruct the audio signal from the stored PSDs to an extent where speech is intelligible and semantic information laid open. To circumvent this, we decided to apply additional smoothing to the PSDs to the point where reconstruction yields no sensitive information.

To determine the required time constant, a listening test was conducted using the Göttingen sentence test (GÖSA; Kollmeier and Wesselkamp, 1997). The speech material was presented to each listener via headphones (HDA200), driven by an amplifier (HB7, Tucker Davis Technologies) and a additional stereo headphone amplifier (MicoAmp HA400, Behringer) adjustable by the test subject. The test was controlled using the Oldenburg Measurement Application (OMA, Hörtech gGmbH). All speech material was presented at a base level of 70 dB SPL without additional backgound noise. For each test condition, the subject was presented one test list and instructed to adjust the volume for best intelligibility using the HA400. After an appropriate level was found, a first list with the unprocessed GÖSA sentences was presented followed by the processed sentences, in randomized order. The sentence lists were also selected randomly. Ten normal-hearing listeners (three female, seven male, age 20-27 years) participated in the test. The respective audiograms showed thresholds of 10 dB HL or below from 125 Hz to 4 kHz and 20 dB HL or below up to 8 kHz.

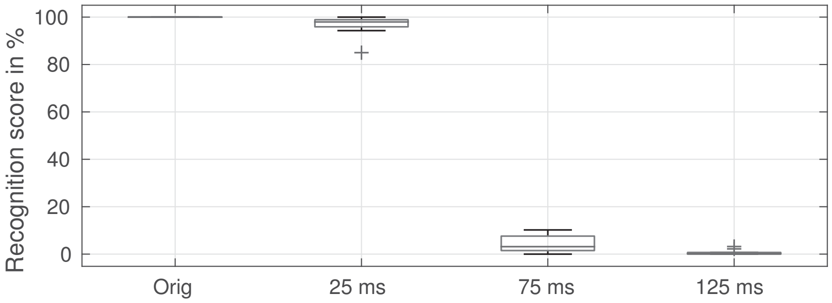

Fig. 4: Recognition score for original and processed sentences ($\tau = 25, 75, \text{and } 125\,\text{ms}$). The boxes show median (boxed line), lower and upper quartile (respective boundary of the box), lowest and highest values within 1.5 times the quartile range relative to the quartiles (whiskers), and outliers (+).

Figure 4 shows the correctly recognized words in percent for each test condition. As expected, the unprocessed sentences were fully recognized. While the median drops slighlty to 97.2% for $\tau = 25$ ms, two listenes still reach a score of 100%. For 75 and 125 ms, the score drops to 1.6% and 0.6% respectively. For the latter, five test listeners could not repeat one single word correctly. While $\tau = 75$ ms also appears to be sufficiently unintelligible, we choose $\tau = 125$ ms for additional headroom in case of uncommon circumstances like exceptionally slow speech.

CONCLUSIONS

This paper describes a well behaved system for long-time analysis of everyday listening situations. It delivers objective acoustic parameters at high resolution while maintaining the privacy of the subject and third parties. The software is easily extendable to capture additional features or provide enhanced functionality. An implementation of a basic online scene-analysis might be used to perform certain actions, e.g., trigger a questionnaire when a the acoustic environment changes, or the user might be prompted to take a picture using the smartphone's camera to capture the scene, of course in accordance with specific privacy regulations.

While the hard- and software are very easy to use, initial field tests showed that elderly people with little or no experience with handheld computers and/or touch-devices sometimes have difficulties operating the system. While there is acoustic and tactile feedback if the audio interface is not plugged in or analysis not started, there is still room to improve the handling as well as instruction of test subjects.

ENDNOTES

[1] The authors are in no way affiliated with the companies mentioned here or have any special interest in promoting a certain product. References are for documentation purposes only.

ACKNOWLEDGEMENTS

Many thanks to Anne Schlüter and Christoph Eike for conducting the listening test and to Audiofon Hearing Systems for supplying the hearing aid dummies.

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

REFERENCES

Kates, J.M. (**2008**). *Digital Hearing Aids*. ISBN: 978-1-59756-317-8. (Plural Publishing).

Kollmeier, B. and Wesselkamp, M. (**1997**). “Development and evaluation of a german sentence test for objective and subjective speech intelligibility assessment,” J. Acoust. Soc. Am., **102**, 2412-1421.

Shiffman, S., Stone, A.A., and Hufford, M.J. (**2008**). “Ecological momentary assessment,” Ann. Rev. Clin. Psychol., **4**, 70-73.

Welch, P.D. (**1997**). “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” IEEE Trans. Audio Electroac., **15**, 70-73.