Development and application of a code system to analyse behaviour in real life listening environments

MARKUS MEIS[1,2,*], MELANIE KRUEGER[1,2], MARIA GEBHARD[1,2,3],
PETRA V. GABLENZ[3,2], INGA HOLUBE[3,2], GISO GRIMM[4,2], AND RICHARD PALUCH[1,2,5]

[1] *Hörzentrum Oldenburg GmbH, Oldenburg, Germany*

[2] *Cluster of Excellence Hearing4all, Oldenburg, Germany*

[3] *Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany*

[4] *HörTech gGmbH, Oldenburg, Germany*

[5] *Carl von Ossietzky University, Oldenburg, Germany*

Numerous studies showed that different hearing aid (HA) algorithms improve speech intelligibility in typical lab situations as measures of clinical efficacy. From the perspective of auditory ecology, it remains obscure to what extent these results really allow for estimating the outcome in listening situations in real life. One promising tool is the observation of participants behaviour induced by different HA settings. We developed an annotation system for coding the behaviour related to the framework of the International Classification of Functioning, Disability and Health (ICF) in iterative steps. The first inputs were derived from a series of lab studies, using virtual acoustics. It was shown that different directional modes of HAs influenced real life behaviour. First indications of activity limitation according to ICF (d3504 'Conversing with many people') were found. Additionally, the behaviour of users in real life was described by means of 'ethnographical walks' outside of the laboratory using field notes. We identified further behaviour patterns addressing spatial awareness. The conversation related ICF sub-categories were validated by analyses of inter-rater reliability (IRR). The outcome of these analyses led to a reformulation of an annotation/coding system for the usage on tablet PCs for instantaneous coding of the test persons behaviour in real life.

INTRODUCTION

In addition to the benefits of hearing aids, as shown in the lab by means of clinical oriented speech tests, hearing specific and generic questionnaires or diaries are used to determine the benefits in every-day life. In several studies it was shown that, without rehabilitation with hearing aids, a hearing loss influences every-day activities negatively and reduces Health-related Quality of Life (HrQoL), particularly in the

*Corresponding author: m.meis@hoerzentrum-oldenburg.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 6: Adaptive Processes in Hearing, August 2017, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C.-Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. The Danavox Jubilee Foundation, 2017.
 ISBN: 978-87-990013-6-1.

domains of social functioning and mental well-being (e.g., Chisolm *et al.,* 2007). HrQoL outcome tools are focusing on self-administered questionnaires over a past period of usually four weeks and are mostly filled in retrospectively, and therefore this can be regarded as a measure of long-term HrQoL (L-HrQoL). Summarizing experiences retrospectively, however, is possibly biased by interlocked effects of memory and subjective perception.

Gatehouse *et al.* (1999) proposed an 'auditory ecology' approach, which takes the objective physical characteristics of every-day listening environments and the individual listener's demands in these real listening environments into account. Up to date our knowledge of real life listening environments still lacks resilient empirical and qualitative data. Possible solutions to bridge this gap could be smartphone-based systems to measure the acoustical environment and to combine those objective measures (acoustical information/data) with subjective data of the respective patient in an approach named Ecological Momentary Assessment (EMA; see, e.g., Bitzer *et al.,* 2016; Kowalk *et al.*, 2017; Shiffman *et al.*, 2008). In addition to subjective ratings of everyday situations over a longer period, short items of "momentary" or acute HrQoL, we called it M-HrQoL, are a promising approach to enrich the outcome toolbox of auditory ecology. M-HrQoL items could be included in measurements of the situation-specific self-perception in the actual situation, as a sub-domain of EMA. Self-perception is still one important data source of listening situations, but behaviour, especially the ability to communicate in conversational situations, is a very relevant outcome area too. Paluch *et al.* (2015) showed that communication behaviour changes in relation to different HA modes. They identified two core dimensions of communication behaviour: 'forms of interaction' (Face-to-Face [F-t-F] vs. group communication) and 'interdependence' (symbolic gestures vs. spoken words) based on Strauss (1987). A higher ratio of F-t-F interactions as well as a higher ratio of verbal communication for an adaptive binaural beamformer in contrast to a broader adaptive monaural beamformer was shown in group conversation, but only in a loud super market scene (L_{Aeq_15min} = 67 dB) in contrast to a softer condition of L_{Aeq_15min} = 55 dB. These behaviour descriptions need to be linked to M-HRQoL to assess the user's handicap qualitatively. The framework of the International Classification of Functioning, Disability and Health (ICF) might facilitate an appropriate approach (WHO, 2001). The ICF model allows to describe the dynamic interaction between the components body functions/structure, activities, participation and environment related, as well as person-centered contextual factors. The ICF model has the privilege to provide generic qualifiers of disability/functioning.

SYSTEMATIC DEVELOPMENT OF A BEHAVIOURAL CODE SYSTEM

Lab test: Comparison of directional modes from three HA devices

In total, six male and four female experienced HA users participated in group discussion sessions (mean age=72.6 years, SD = 7.6, $PTA_{4\ (0.5,\ 1.0,\ 2.0,\ 4.0\ kHz)}$ better ear = 49.7 dB HL, SD = 6.7). The participants were divided into two groups of five subjects, which were invited successively. In the experiment, the participants were seated at one table with near and distant communication partners for four group

sessions with a duration of 15 minutes each. For the study, three custom-made in-the-ear (ITE) devices from different brands were fitted (first fit) to the test persons (for preliminary data see Latzel *et al*., 2016). Three devices per ear were built from the identical ear impression for each subject. The power levels of the hearing aids were specified in order to compare the same power levels across test devices. In each HA, a program for speech intelligibility in loud situations was fitted with a directional microphone mode with narrow directionality. The vents of the hearing devices were individually chosen due to the pure tone audiogram and the HA characteristics. The realization of the group discussion procedure was exactly the same as in the study from Paluch *et al.* (2015), but took place only in a noisy scene (L_{Aeq_15min} = 67 dB). After each of the four conversation sessions, a questionnaire was filled out by the participants. For the subjective rating of speech intelligibility a scale from '1' (nothing) to '7' (all) was used. For the analyses of the behavioural data the same annotation scheme as in the Paluch *et al.* (2015) study was applied.

At first glance, the data showed that no differences according the dimensions F-t-F vs. group communication and symbolic gestures vs. spoken words were observed. This pattern of results was contradictory to the subjective ratings of the participants regarding perceived speech intelligibility. Further analysis established statistically significant differences (non-parametrical analyses of repeated measurements) in speech intelligibility ratings for the three devices, indicating that, e.g., device #3 was rated with a median of 2 and device #1 with a median of 3.5. The obtained differences did not reflect the behavioural data; Therefore, a clarification was necessary. A team of three raters inspected once again the whole video material. In an iterative Grounded Theory (Glaser and Strauss, 1967) based process, two further sub-dimensions were striking: different proxemics regarding near vs. distant torso movements (forward-backward) to the dialogue partner and conversations with the distant vs. near dialogue partner. We proposed thus a revised annotation scheme (Meis *et al.*, 2016), as illustrated in Fig. 1.

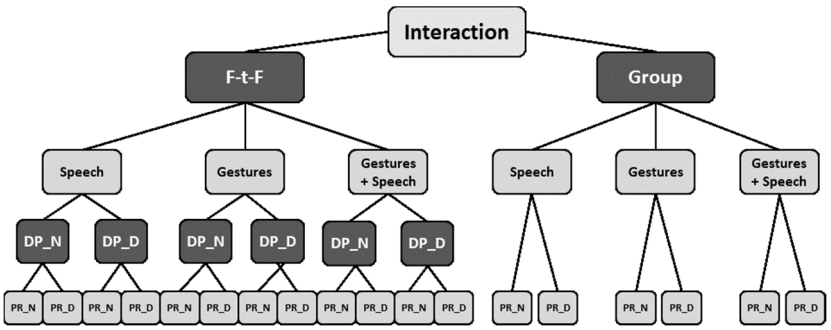

Fig. 1: 18 code annotation scheme of communication / interaction. Face-to-Face: F-t-F Interaction, Group: Group Interaction, DP: Distance Partner: near vs. distant, PR: Proxemics: near vs. distant.

The result was an annotation scheme including in total 18 codes, a hierarchic scheme of interdependent codes, for the four behaviour domains. Using this scheme, 2,939 behaviour units with a time resolution of ~13 s per unit/test person were assessed.

Following the revised annotation scheme, no differences were found regarding the core dimension 'interdependence'. Regarding the core dimension 'forms of interaction' the ratio of F-t-F and F-t-F plus group interaction was highest for device #3, indicating nearly 15% more interactions in contrast to the two other devices. The examination of the F-t-F category 'Distance to the dialogue partner' (near vs. distant) showed that test persons using device #3 interacted in > 80% (median) of the assessed interaction units only with the respective near dialogue partner, in contrast to device #1 (Wilcoxon signed rank test, p=0.009). Using device #1, the ratio of interactions was balanced regarding the F-t-F communication of the near vs. distant dialogue partner as shown in Fig. 2. Additionally, the analysis of proxemics revealed the significant effect that subjects tended to lean more forward for device #2 (p=0.043, Wilcoxon signed rank test) and #3, compared to subjects using device #1.

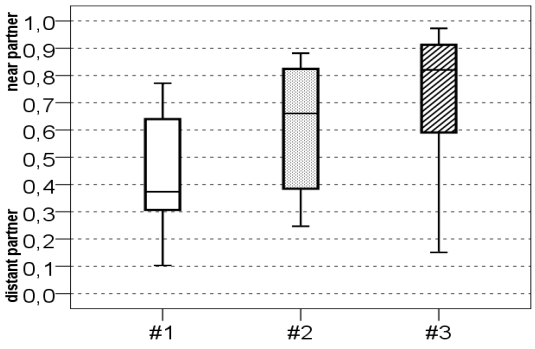

Fig. 2: Communication with near vs. distant dialogue partner for three different devices in %. Boxplots show the distribution of the ratios calculated from F-t-F near partner/near and distant partner communication.

The data regarding 'Distance to the dialogue partner' can be interpreted as a limitation of communication activities induced by the microphone mode of the hearing aids. Subjects with sub-optimal HA fittings are rather able to communicate with the near dialogue partner, but not with the respective distant dialogue partner. This result pattern suggested the interpretation of data along theoretical and practical models of HrQoL and the classification of the ICF framework.

Expert review and ethnography

Granberg *et al*. (2014) published a comprehensive ICF core set for hearing loss with the domains 'body functions' ('b' codes), 'body structures' ('s' codes), 'activities and participation' ('d codes'), and 'environmental factors' ('e codes') with in total over 100 codes. This Core Set was used as a basis for a review meeting with six experts in the field of audiology, rehabilitation, and psychology with the main goal to extract

codes applicable in *behavior observations* using a 3-point scale. The most prominent ICF codes for behavior analyses were derived from the functions *'activities and participation'*, labeled as 'd' codes and – partly related – 'b' codes.

Codes rated 'moderate appropriate' or 'very appropriate' by the experts were used for external observations in an ethnographic field study (Paluch *et al*., 2017). This ethnographic study was conducted as a stand-alone experiment with 10 test persons (n=2 normal-hearing, n=7 unaided slight to moderate hearing loss, and n=1 moderate hearing loss aided with HA; reference better ear PTA_4 according to WHO, 2001). Three test persons classified as unaided were externally observed both unaided and recently fitted with hearing aids. The behavior of the other seven participants was observed in the respective aided or unaided condition. During the external observation the subjects took a 4.5-km walk and visited different locations (cafeteria, several bus stops). A trained observer generated field-notes, based on the methodology of Przyborski and Wohlrab-Sahr (2009). Paluch *et al*. (2017) showed that newly fitted hearing-impaired subjects tended to move their torsos and heads (left-right level) in a significant manner, possibly caused by new auditory input, particularly spatial input.

Based on the results of the reported studies, the expert review, and the ethnographical walks an extended set of ICF categories for behaviour analyses in the field was finally derived (Table 1).

Inter-rater reliability (IRR) of the extended ICF core sets

The future goal is to develop a tool to assess instantaneous behaviour ratings in the field via tablet PC and to combine these external observed behavioural data with objective and subjective data for a multifaceted EMA. Therefore, the proposed extended ICF categories needed a check by IRR procedures, addressing – in a first step – conversation situations.

For the IRR check, a manual with the detailed descriptions of the extended ICF categories was elaborated to provide a clear reference for the evaluation of the different characteristics. IRR was established by three different raters.

- **d160** Focusing attention
    1. Movements torso horizontal axis, frequency <= 45° vs. >45°
    2. Movements head, horizontal axis, frequency <= 45° vs. >45°
- **b140** Attention functions
    1. Sustained attention: face of conversation partner, strong vs. weak
- **d3504** with many people/ **d3503** Conversing with one person
    1. F-t-F vs. group (**only d3504**)
    2. Frequency general verbal communication
    3. Communication partner: distant vs. near (only **d3504**)
    4. Proxemics (torso position) lean forward/backward vertical axis 90°
    5. Frequency change sitting position
    6. Non-understanding gestures, frequency
    7. Speech supporting gestures, frequency

Table 1: Extended ICF categories for behaviour ratings in the field.

Rater / ICF (sub-) categories/scale	A-B		B-C		A-C	
	κ	r_{Sp}	κ	r_{Sp}	κ	r_{Sp}
b140_1 Sustained attention face partner: low-medium-high	.39	.58	.32	.56	**.44**	.65
d3504_1 Communication: F-t-F-balanced-group	**.47**	.58	.36	.38	**.57**	.70
d3504_2 Frequency verbal comm.: seldom-sometimes-frequent	**.51**	.72	**.52**	.68	**.43**	.70
d3504_3 Communication partner: near-balanced-distant	**.59**	.73	**.62**	.70	**.72**	.79
d3504_4 Proxemics: forward-balanced-backward	**.57**	.68	.38	.52	**.50**	.59
d3504_5 Change torso position: seldom-sometimes-frequent	.13	.26	.33	.56	.39	.57
d3504_6 Non-understanding gestures: seldom-sometimes-frequent	.07	.29	.35	.40	.16	.32
d3504_7 Speech supporting gestures: seldom-sometimes-frequent	.24	.51	.26	.39	**.46**	.57

Table 2: IRR for extended ICF categories. Cohen's kappa indicating moderate or substantial agreement in bold. A-C = 3 raters; κ = Cohen's kappa; rSp = Spearman's rho. Cohen's kappa agreement: <0 = "poor", 0–0.20 = "slight", 0,21–0,40 = "fair", 0,41–0,60 = "moderate", 0,61–0,80 = "substantial", 0,81–1,00 = "almost perfect"; see Landis and Koch (1977).

The raters had to rate a selection of two video sessions of the ITE benchmark study, presented above. The video material included five subjects and two different devices. In contrast to the study from Latzel *et al*. (2016), the rating referred to 3-min sections (in total 5 ratings in a 15-min conversation) with 5- or 7-point rating scales in order to reduce too frequent annotation activities for the rater in a field situation. We calculated Cohen's Kappa (κ) (Cohen, 1960) and correlations (Spearman's rho r_{Sp}) for pairs of raters to get deeper insight of the rater characteristics and condensed the rating scales into 3-point ordinal scales; see Table 2.

We observed predominantly poor to slight IRR statistical values for the categories b140_1 and d3504_5 to d3504_7. Moderate IRR-values were assessed for the categories F-t-F vs. group (d3504_1), frequency verbal communication (d3504_2), and proxemics (torso position) lean forward/backward vertical axis (d3504_4). Substantial IRR-values were gathered for the interactions with the distant vs. near conversation partner (d3504_3).

It is planned to use category 'd160' only for spatial awareness topics with moving sources, which are not included in the external communication behaviour assessment. Therefore, this category was not included in the IRR procedure.

DISCUSSION AND OUTLOOK

The development and application of a code system to analyse behaviour in real life listening environments was outlined in this paper. Based on the first explorative studies it was shown that ICF categories are related significantly with hearing aid usage, especially signalling *activity limitation* and *participation restriction*. In

addition to clinical outcome measures, behavioural data of complex interaction episodes of group conversations in noisy environments offer the possibility to use auditory ecological valid outcome measures, which capture how hearing aids impact behaviour in real life. The approach and the studies presented here should be understood as explorative. They certainly need further theoretical foundation as well as the proof of reproducibility. The IRR-values indicated moderate to substantial inter-rater agreement in relevant ICF categories, but the IRR has to be improved for the usage in the field. The three raters stated that they had difficulties to average different behaviour units inside a three minute section. Moreover, it might be easier to annotate the conversation behaviour directly, e.g., on a tablet PC with a graphical user interface (GUI) for quick and easy tapping, but with a reduced set of codes. In future, we propose to use six hierarchic and interdependent main codes to evaluate group conversation, which include the categories 'F-t-F vs. group', 'near vs. distant dialogue partner', and 'near vs. distant proxemics' plus two codes of non-verbal proxemics 'near vs. distant proxemics' during listening. Using instantaneous annotations, the frequency of verbal communication episodes automatically will be recorded. The GUI should be completed with ICF relevant environmental and contextual categories, such as light condition (e240). In the next studies, we are going to combine the proposed eight codes of external behavioural observation with objective acoustical data and subjective items to get a more complete picture of hearing impaired users in real listening environments. In future, the approach reported here, should be combined and/or validated with the automatic assessment of behavioural data, such as head- and eye-tracking procedures.

ACKNOWLEDGMENTS

This work was supported by the Cluster of Excellence EXC 1077/1 *Hearing4all,* funded by the German Research Council (DFG), Matthias Latzel from Sonova, and the Hearing Industry Research Consortium (IRC) 2016 grant.

REFERENCES

Bitzer J., Kissner S., and Holube I. (**2016**). "Privacy-aware acoustic assessments of everyday life," J. Audio Eng. Soc., **64**, 395-404.

Chisolm T.H., Johnson C.E., Danhauer J.L., Portz L.J.P., Abrams H.B., Lesner S., McCarthy P.A., and Newman C.W. (**2007**). "A systematic review of health-related quality of life and hearing aids: Final report of the American Academy of Audiology task force on the health-related quality of life benefits of amplification in adults," J. Am. Acad. Audiol., **18**, 151-183.

Cohen, J. (**1960**). "A coefficient of agreement for nominal scales," Educ. Psychol. Meas., **20**, 37-46.

Gatehouse S., Elberling C., and Naylor G. (**1999**) "Aspects of auditory ecology and psychoacoustic function as determinants of benefits from and candidature for non-linear processing in hearing aids," Proc. Danavox Symposium, **18**, 221-233.

Glaser, B.G., and Strauss, A.L. (**1967**): *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago: Aldine.

Granberg, S., Möller, K., Skagerstrand, A., Möller, C., and Danermark, B. (**2014**). "The ICF Core Sets for hearing loss: researcher perspective, Part II: Linking outcome measures to the International Classification of Functioning, Disability and Health (ICF)," Int. J. Audiol., **53**, 77-87. doi: 10.3109/14992027.2013.858279.

Kowalk, U., Kissner, S., v. Gablenz, P., Holube, P., and Bitzer, J. (**2017**). "An improved privacy-aware system for objective and subjective ecological momentary assessment," Proc. ISAAR, **6**, 25-30.

Landis, J.R., and Koch, G.G. (**1977**). "The measurement of observer agreement for categorical data," Biometrics, **33**, 159-174.

Latzel, M., Paluch, R., Meis, M., and Krueger, M. (**2016**). **"**A new tool for subjective assessment of hearing aid performance: Analyses of interpersonal communication—next step(s)," Poster B12, International Hearing Aid Research conference (IHCON), Tahoe City, CA.

Meis, M., Paluch, R., Krueger, M., and Latzel, M. (**2016**). **"**A new evaluation tool for hearing aids in everyday situations: Video-based analysis of interpersonal communication behavior, part 2," 61st International Congress of Hearing Aid Acousticians, Hannover.

Paluch R., Latzel, M., and Meis, M. (**2015**). "A new tool for subjective assessment of hearing aid performance: Analyses of interpersonal communication" Proc. ISAAR, **5**, 453-460.

Paluch, R., Krueger, M., Grimm, G., and Meis, M. (**2017**)**.** "Moving from the field to the lab: Towards ecological validity of audio-visual simulations in the laboratory to meet individual behavior patterns and preferences," 20. Jahrestagung der Deutschen Gesellschaft für Audiologie, Aalen.

Przyborski, A., and Wohlrab-Sahr, M. (**2009**). *Qualitative Sozialforschung. Ein Arbeitsbuch. 2., korrigierte Auflage.* München: Oldenburg Verlag, 403 Seiten, 978-3-486-59103-3.

Shiffman, S., Stone, A.a., and Hufford, M.R. (**2008**). "Ecological momentary assessment," Ann. Rev. Clin. Psychol., **4**, 1-32. doi: 10.1146/annurev.clinpsy.3.022806.091415

Strauss, A.L. (**1987**). *Qualitative Analysis for Social Scientists.* New York: Cambridge University Press.

WHO (**2001**). *International Classification of Functioning, Disability and Health: ICF.* Geneva: World Health Organization.