Testing a Quality of Experience (QoE) Model of Loudspeaker-based Speech Reproduction

Stefan Uhrig[1,2], Sebastian Möller[1,3], Dawn M. Behne[4], Peter Svensson[2], Andrew Perkis[2]
[1]Quality and Usability Lab, Technische Universität Berlin, Germany
[2]Department of Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway
[3]Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[4]Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway
stefan.uhrig@qu.tu-berlin.de, sebastian.moeller@tu-berlin.de, {dawn.behne, peter.svensson, andrew.perkis}@ntnu.no

***Abstract*—This study introduces a Quality of Experience (QoE) model of loudspeaker-based speech reproduction, which specifies quality elements and quality features relevant to Overall Listening Experience (OLE) and Quality of Service (QoS), respectively. Assumptions about the relations between selected quality elements and quality features were validated in a listening-only test. Participants had the task to behaviorally identify the voices of two different talkers. The talkers took turns in uttering sentences through only a central loudspeaker (non-spatial mode) versus through either the central or one talker-specific lateral loudspeaker (spatial mode). The quality of the transmitted speech signals was either clean, superimposed with background noise or bandpass-filtered. It was demonstrated that *transmission quality*, but not *reproduction mode* significantly influenced evaluative (speech quality, speech intelligibility) and immersive (voice naturalness, spatial presence, social presence) aspects of listening experience. Unexpectedly, the spatial mode did not reduce the mental effort of talker identification, as opposed to prior evidence. The results suggest that noticeable advantages of spatial hearing in speech reproduction only manifest in listening situations of higher complexity. Moreover, the employed subjective measures (category rating scales) might not have been sensitive enough to capture more subtle variation in behavioral task performance.**

I. Introduction

Challenges for speech perception arise in situations when multiple talkers are active, especially during conversations (e.g., conference calls via Skype). The human auditory system possesses the ability to localize distinct talker sound sources with high sensitivity [1], which allows to segregate different sound streams and facilitate subsequent perceptual and cognitive processing. It achieves this through rapid pick-up of spatial auditory cues from incoming sound streams [2]. Since most conventional speech communication technologies mix all talkers' voices together through mono-channel sound reproduction, spatial cues inherent to natural acoustic environments are lost. Yet, newly emerging multimedia technologies like auditory virtual and augmented reality have regularly been implementing spatial sound reproduction techniques to enable users to more effectively and efficiently process information originating from physical/virtual sound sources [3], [4].

In particular, audio tele-conferencing has been targeted as a new multimedia service whose functionality would most likely benefit from spatial speech reproduction [5]. Typical tasks performed via tele-conferencing technologies in multi-talker listening situations require identification of different talkers. An obvious advantage of spatial compared to non-spatial audio tele-conferencing lies in its improved talker localization. Besides, spatialized tele-conferences might heighten the sense of being part of a conversational scene [6] and evoke feelings of "togetherness" with the other talkers [7].

The acceptance and successful adoption of new multimedia technologies, including their utilization of spatial sound, critically depends on the *quality* of respective systems, services and applications as experienced by human users [8]. A common approach towards quality assessment and evaluation establishes relations between technical properties and perceptual attributes that contribute to overall quality experience, referred to as *quality elements* and *quality features*, respectively [9], [10]. However, due to the highly immersive and interactive character of these new technologies, other higher-order perceptual attributes besides quality as well as task-related and contextual influencing factors must be taken into account [5].

This study proposes a Quality of Experience (QoE) model to explain dependencies between selected quality features and quality elements relevant to the reproduction of transmitted speech signals via physical loudspeakers (see next Sec. II). To evaluate the impact of a subset of quality features in the context of talker identification, results of a listening test are presented and interpreted with respect to prior work (Sec. III).

II. Quality of Experience (QoE) Model

Rumsey put forward a "scene-based paradigm" for defining perceptual (spatial and timbral) attributes of spatially reproduced auditory scenes [11]. Accordingly, auditory scenes should be static, but allow for grouping of (micro and macro) scene elements, which might include single sound sources, groups of sound sources and the background environment. In addition, influences of the behavioral task and context are considered to be crucial. Different authors have proposed taxonomies for perceptual attributes of auditory scenes, some of which might also serve as quality features [12], [13]:

Dimensional [11], *physical* [14] or *geometrical* [6] attributes are exclusively bottom-up influenced by parameters of the physical/virtual environment and objects/events within it, constituting the scene and its elements, respectively. Examples include perceived height, width and depth of sound sources

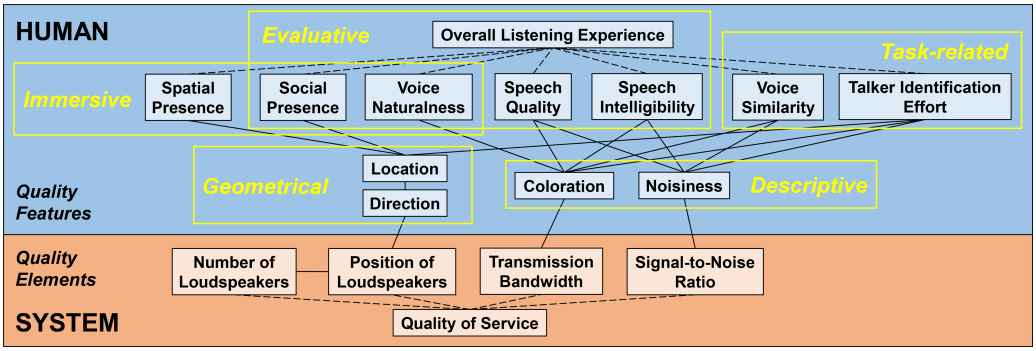

Figure 1: Quality of Experience (QoE) model of loudspeaker-based speech reproduction in the context of a talker identification task. Subjective and objective domains of human users' perception (blue area) and technical system (red area) are distinguished, respectively. The human domain contains perceptual attributes contributing to *Overall Listening Experience* (quality features), which are grouped into categories (geometrical, descriptive, immersive, evaluative, task-related). The system domain includes technical properties affecting quality features (quality elements), the totality of which are subsumed under *Quality of Service*.

[11], [12]; also, perceived direction and distance of sound sources relative to listener position (*perceived location*) [12].

Immersive [11], *psychic and affective* [14] or *general impression* [6] attributes are strongly top-down influenced by expectations about a listener's own bodily self, emotional and motivational state as well as prior knowledge. Examples include the perception of sound sources as "outside-the-head" (*externalization*) [6], [12], [15]; the sense of being surrounded by sound sources or a reverberant sound field (*envelopment*) [4], [6]; the sense of self-location within an auditory scene (*spatial presence*) [6], [11]; the feeling of being together with other sentient entities in the scene (*social presence*; also: "co-presence") [7]; and the congruence/coherence of scene elements to/with previously experienced or imagined equivalent ones (*naturalness*, *authenticity*, *plausibility*) [6], [14], [16], [17].

Following the development of the Quality of Experience (QoE) approach, a deeper understanding of "quality" emerged, emphasizing its affective character, the user's perspective (e.g., personality traits, emotional/motivational state) and multi-layered contextual influencing factors [8]. Accordingly, quality is defined as a higher-order evaluative perceptual attribute, whereas lower-order perceptual attributes that map onto quality (or "preference") are regarded as descriptive [11], [18]. In the audio/speech domain, the QoE approach is reflected in the concept of *Overall Listening Experience* [14], [19]; three orthogonal quality features (dimensions) of speech transmission quality, "discontinuity", "noisiness" and "coloration", have been linked to quality elements like packet/frame loss rate, signal-to-noise ratio and transmission bandwidth, respectively, in case a forth quality feature "loudness" is kept constant [20].

Figure 1 illustrates a QoE model for loudspeaker-based speech reproduction in the context of a human talker identification task, which proposes five, partially overlapping categories of perceptual attributes (potentially serving as quality features): Geometrical, descriptive, immersive, evaluative and task-related. The QoE model specifies presumed relations between relevant quality features and quality elements manipulated by the listening test to be described in Section V.

III. Related Work

Past research has revealed effects of spatial speech reproduction on subjective measures of preference, quality and effort:

Baldis et al. conducted a listening-only test with several conversation scenarios [21]. Spatialization was achieved by playback over a single loudspeaker (non-spatial condition) or four loudspeakers positioned in a semicircular array (spatial). The test indicated a higher preference as well as reduced experienced difficulty and mental effort regarding talker identification for spatial versus non-spatial sound reproduction.

In another study by Kilgore et al. [22], participants listened to conversation scenarios either through mono format (non-spatial) or stereo format including binaural location cues (spatial). Participants again preferred spatial over non-spatial sound reproduction, experienced lower difficulty and reduced mental effort to identify talkers.

In initial studies by Raake et al., participants either attentively listened to pre-recorded conversations (listening-only test) or engaged in conversations with the other interlocutors (conversation test), with sound reproduction mode being either spatial or non-spatial [23], [24]. For both listening-only and conversation tests, preference and quality were rated higher in the spatial versus non-spatial mode and mental effort was reduced while listening to (listening effort) or actively taking part in the conversation (conversation effort). For listening-only, spatialization also enhanced speech intelligibility and talker recognition, referring to the experienced ability to understand semantic speech content and recognize interlocutors, respectively. Also, the spatial mode was rated as more useful.

Later on, Skowronek and Raake conducted listening-only tests using multi-party conversation scenarios for audio teleconferencing in order to examine effects on several subjective measures of quality and mental effort [25], [26]. It turned out that dichotic (spatial) versus diotic (non-spatial) sound reproduction was increasing quality (overall quality, connection quality) and speech intelligibility as well as reducing mental effort (e.g., concentration effort, talker identi-

fication/recognition effort, topic comprehension effort). Some effects of spatialization on mental effort seemed to be more pronounced for higher numbers of interlocutors, that is, listening conditions with higher baseline levels of mental effort due to increased difficulty in perceptual separation of talkers.

IV. Listening Test: Expectations

The present study pursues to investigate effects of spatial versus non-spatial speech reproduction and transmission quality impairments on subjective experience in a listening-only situation with two talkers. Different types of noise (e.g., background noise, signal-correlated noise) and bandwidth limitations (e.g., bandpass-filtering) are well-known to affect perceived quality and intelligibility of transmitted speech [18], [20], [24], [27]. Usually, spectral distortions (incl. bandwidth limitations) are also inversely correlated with voice naturalness [27], [28]. From a task performance viewpoint, quality degradations should increase voice similarity and in turn heighten the mental effort involved in talker identification [24]–[26]. Due to the more scenic character of spatial speech reproduction—involving sound sources at varying loudspeaker locations—participants would presumably hold stronger impressions of spatial presence and more easily develop feelings of social presence. Lastly, spatial speech reproduction should lower talker identification effort [24]–[26]. The above-mentioned relationships can be construed as links between quality features and quality elements specified in the QoE model shown in Figure 1.

V. Methods

A. Participants

Subjective data were collected from $N = 32$ participants (age: $M = 26.8$, $SD = 5.9$, $R = 14 - 44$ years; 11 female; 5 left-handed). All of them were native Norwegian speakers with normal or corrected-to-normal vision and normal hearing. Each participant received a cinema ticket as compensation.

B. Stimuli

40 phonetically rich Norwegian sentences of varying duration ($M = 4.9$, $SD = 1.5$, $R = 2.1 - 8$ s), uttered by two male native Norwegian speakers in the Oslo dialect, served as stimulus material. The speech recordings were taken from the public "NB Tale basic acoustic phonetic speech database for Norwegian", offered by the National Library of Norway.[1] The sentences had been manuscript-read and concerned arbitrary, neutral topics. For each talker, 20 different sentences were available: The semantic content of 3 sentences was the same for both talkers, while the content of the remaining 17 sentences was specific to each talker. The talkers' voices were unknown to the participants prior to the experiment.

The clean source files of all 40 stimuli were degraded using functions from the P.TCA toolbox for MATLAB software (version R2018a) [29]: Addition of stationary pink noise, targeting -5 dB signal-to-noise ratio, produced 40 noisy stimuli;

[1] https://www.nb.no/sprakbanken/show?serial=sbr-31

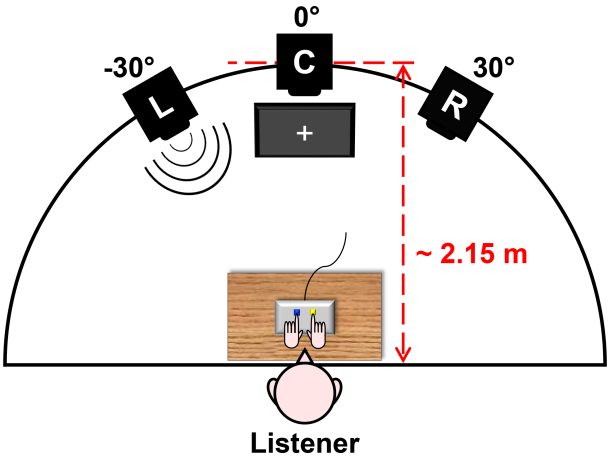

Figure 2: Test layout used in the present study. The listener is sitting at a table in front of three loudspeakers on left (L), center (C) and right (R) locations with equal angles (L = -30°, C = 0°, R = 30° azimuth) at a distance of approximately 2.15 m. During audio playback, the listener is fixating a white cross on the monitor under the central loudspeaker.

Application of a bandpass Butterworth filter with a low-cutoff frequency of 400 Hz and a high-cutoff frequency of 800 Hz produced 40 filtered stimuli.

In a final step, the total number of 40 clean plus 80 degraded (noisy, filtered) stimuli were normalized to -26 dBov active speech level.

C. Experimental Procedure

All testing sessions were carried out in a quiet, sound-attenuated laboratory room, lasting approximately one hour.

The participants were seated at a small table facing a semi-circular array of three, equiangularly separated loudspeakers (Dynaudio BM6A) mounted on stands approximately at height of the listener's head as illustrated in Figure 2. An elevated standard computer monitor was positioned on the floor below the central loudspeaker.

A repeated-measures experimental design with two fully crossed factors, *reproduction mode* (non-spatial, spatial) and *transmission quality* (clean, noisy, filtered), resulted in 6 conditions (non-spatial/clean, non-spatial/noisy, non-spatial/filtered, spatial/clean, spatial/noisy, spatial/filtered). The experiment consisted of 6 test blocks for every experimental condition. In the non-spatial mode, both talkers were presented through the central loudspeaker; in the spatial mode, half of the trials of each talker were presented through one talker-specific lateral loudspeaker and the other half through the central loudspeaker.

During each block, all 40 stimuli from the current condition were serially presented through the loudspeaker(s) with an inter-stimulus interval (ISI) of 1.5 s, randomly jittered by $\pm$ 0.5 s. Thus, each sentence had finished before the next one started after the ISI, reflecting a turn-taking scenario without any simultaneous talk. The order of blocks (i.e., experimental conditions) was randomized across participants; the order of stimuli (i.e., sentences and talkers) was pseudo-randomized

across blocks and participants, such that each stimulus was followed by a stimulus with different semantic content. Stimulus presentation was controlled by Psychophysics Toolbox Version 3 (PTB-3)[2] for MATLAB. A high-quality audio interface (Roland UA-1610 Studio-Capture) was used for audio playback. Master volume was set to a comfortable listening level around 65 dB at the listener position.

During stimulus presentation, the participants' behavioral task was to quickly identify the talker after each new stimulus had started by pressing buttons on a response pad. They were instructed to fixate a white cross on the monitor to keep their head position constant.

After stimulus presentation ended, a series of 7 category rating scales was presented on the monitor screen. All scales were continuous and extended at the extremities according to [30], with 7 major scale points and 4 minor scale points in-between two major ones (see Fig. 3).[3] Descriptive labels were attached to the major scale points. Participants used a computer mouse to move a cursor along the scale and select a convenient position by left-clicking, after which the next scale would appear. They had read details on the meaning and proper usage of the scales in the task instructions. The order of scales was randomized across blocks and participants.

D. Data Analysis

Repeated-measures analyses of variance (ANOVAs) were computed by use of the "ez"[4] package for R. In total, 7 independent ANOVAs with *reproduction mode* (non-spatial, spatial) and *transmission quality* (clean, noisy, filtered) as within-subject factors were fitted to each category rating (speech quality, speech intelligibility, voice similarity, voice naturalness, spatial presence, social presence, talker identification effort; see Fig. 3) as dependent variable. A statistical significance level of $\alpha = 0.05$ was chosen and Šidák-adjusted for the 7 ANOVAs. Generalized eta squared (η^2_{G}) was computed as an effect size measure. For post-hoc comparisons, paired t-tests with Holm correction were calculated.

In addition, a correlational analysis was conducted to estimate 21 Pearson correlation coefficients (r) and p-values for each pair of the 7 subjective measures (across all conditions), adjusted for multiple testing with the Holm correction.

VI. Results

Figure 4 contains mean plots for effects of *reproduction mode* (non-spatial, spatial) and *transmission quality* (clean, noisy, filtered) on each subjective measure.

Analysis of subjective ratings yielded statistically significant main effects of *transmission quality* on speech quality ($F[2, 62] = 357.69$, $p < 0.001$, $\eta^2_{\mathrm{G}} = 0.83$), speech intelligibility ($F[2, 62] = 114.89$, $p < 0.001$, $\eta^2_{\mathrm{G}} = 0.67$), voice similarity ($F[2, 62] = 21.39$, $p < 0.001$, $\eta^2_{\mathrm{G}} = 0.11$), voice naturalness ($F[2, 62] = 63.61$, $p < 0.001$, $\eta^2_{\mathrm{G}} = 0.48$), spatial presence ($F[2, 62] = 53.40$, $p < 0.001$, $\eta^2_{\mathrm{G}} = 0.37$), social presence ($F[2, 62] = 52.83$, $p < 0.001$, $\eta^2_{\mathrm{G}} = 0.37$) and talker identification effort ($F[2, 62] = 32.54$, $p < 0.001$, $\eta^2_{\mathrm{G}} = 0.18$).

Post-hoc pairwise comparisons between clean and degraded (noisy, filtered) stimuli were significant (all $p < 0.001$); comparisons between noisy and filtered stimuli were significant for speech quality ($p < 0.001$), speech intelligibility ($p < 0.001$), voice naturalness ($p < 0.01$) and talker identification effort ($p < 0.01$).

Neither main effects of *reproduction mode* nor any interaction effects were significant.

Results from the correlational analysis are listed in Table I, with rows sorted by r in decreasing order.

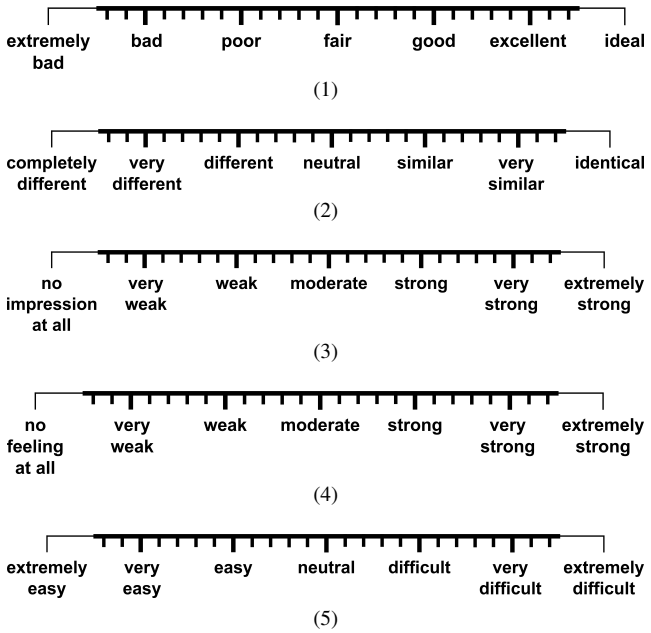

Figure 3: Category rating scales used for assessing 7 subjective constructs, as operationalized by the following questions:

- **Speech quality**: *How was the quality of the speech transmission?* (1)
- **Speech intelligibility**: *How was the intelligibility of the speech transmission?* (1)
- **Voice similarity**: *How similar did the transmitted voices of the speakers sound?* (2)
- **Voice naturalness**: *How similar did the transmitted voices sound to natural voices?* (2)
- **Spatial presence**: *How strong was your impression of being in a room with the speakers?* (3)
- **Social presence**: *How strong was your feeling of being together with the speakers?* (4)
- **Talker identification effort**: *How difficult was it for you to identify the speakers?* (5)

[2] http://psychtoolbox.org/

[3] This "extended mean opinion score" scale design counteracts biases of cognitive judgment in category ratings [27]. In regard to immersive attributes of spatial and social presence, it is noted that participants could report zero intensities by selecting extreme left labels ("no impression/feeling at all").

[4] https://cran.r-project.org/web/packages/ez/

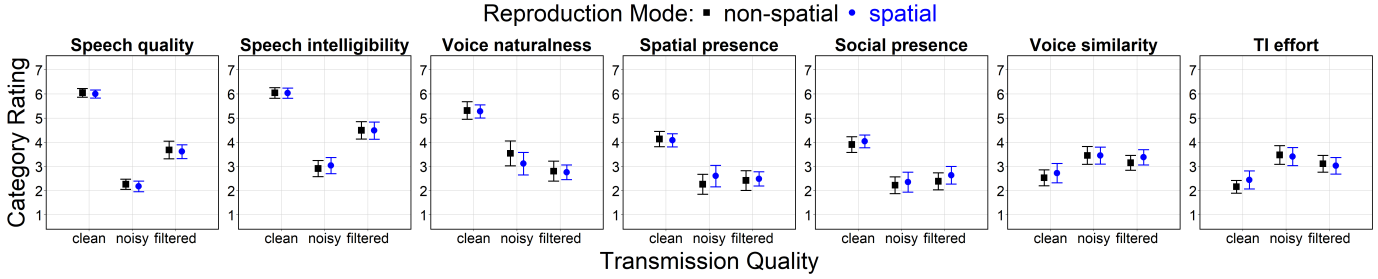

Figure 4: Effects of *reproduction mode* and *transmission quality* on category rating for each subjective measure ($N = 32$). Scale numbers correspond with scale point labels depicted in Figure 3. Error bars represent 95% confidence intervals.

Table I: Pearson correlation coefficients r with 95% confidence intervals [LL = lower limit, UL = upper limit] for each pair of scales. All r were statistically significant with $p < 0.001$.

Scale 1	*Scale 2*	r $[LL, UL]$
Speech quality	Speech intelligibility	0.85 $[0.78, 0.90]$
Spatial presence	Social presence	0.84 $[0.76, 0.89]$
Voice similarity	TI effort	0.68 $[0.54, 0.78]$
Speech quality	Voice naturalness	0.62 $[0.47, 0.74]$
Speech quality	Social presence	0.60 $[0.44, 0.72]$
Speech quality	Spatial presence	0.58 $[0.42, 0.71]$
Speech intelligibility	Voice naturalness	0.55 $[0.38, 0.68]$
Voice naturalness	Spatial presence	0.54 $[0.37, 0.67]$
Speech intelligibility	Social presence	0.53 $[0.36, 0.66]$
Voice naturalness	Social presence	0.53 $[0.36, 0.66]$
Speech intelligibility	Spatial presence	0.51 $[0.34, 0.65]$
Voice similarity	Spatial presence	-0.24 $[-0.37, -0.10]$
Voice similarity	Social presence	-0.31 $[-0.45, -0.15]$
Speech intelligibility	Voice similarity	-0.35 $[-0.49, -0.19]$
Spatial presence	TI effort	-0.37 $[-0.52, -0.21]$
Speech quality	Voice similarity	-0.41 $[-0.55, -0.24]$
Social presence	TI effort	-0.44 $[-0.58, -0.28]$
Voice naturalness	TI effort	-0.46 $[-0.60, -0.29]$
Speech intelligibility	TI effort	-0.46 $[-0.60, -0.29]$
Voice similarity	Voice naturalness	-0.47 $[-0.61, -0.30]$
Speech quality	TI effort	-0.50 $[-0.64, -0.33]$

VII. Discussion

The results from the listening-only test confirmed the impact of *transmission quality* on speech quality and intelligibility, with stronger effects being found for noisy versus filtered speech similar to earlier result patterns [18]. The closest correlational relationship was established between speech quality and speech intelligibility ($r = 0.85$, see Tab. I), which pointed towards a close interdependence between them.

Vice versa, perceived voice naturalness was reduced more strongly for filtered speech than noisy speech as was to be anticipated from previous work by Moore and Tan [28]; possibly, participants were to some degree able to discern the intact speech signal from the superimposed background noise. Here, the term "naturalness" exclusively denotes lower-level expectations regarding the physical surface form of the reproduced speech signal, but not its meaning (semantic content) as well as functional significance [9], [17]. Thus, it constitutes an aspect of the broader concept of *plausibility* [6], [16], comprising also expectations at higher levels of abstraction, whose experimental manipulation would require semantic/functional violations in relevant content dimensions.

The induced speech quality impairments either masked (noisy) or eliminated spectral portions of (filtered) individual voice characteristics, hereby making the two voices sound more similar which in turn increased the experienced difficulty of talker identification ($r = 0.68$). In addition, spatial presence and social presence were reduced by degraded speech transmissions of both kinds. It might be argued that during the clean listening condition a "moderate" impression/feeling of presence had been evoked, since voices were sounding very similar to a situation as if talkers were actually talking to the listener in the room. In degraded conditions, however, the induced background noise and filtered voice characteristics would give away the mediated nature of the speech transmission and break any experienced presence. The two subjective constructs were further highly correlated ($r = 0.84$), suggesting that both types of presence overlapped considerably.

Surprisingly, quality-dependent differences in experienced presence manifested although none of the effects of *reproduction mode* turned out to be significant. The availability of spatial auditory cues in the spatial reproduction mode did not shift subjective ratings towards higher presence. Presenting speech stimuli via a binaural headphone system, a recent study by Werner et al. demonstrated that the spatial complexity of auditory scenes influenced subjective ratings of spatial presence [31]. Presumably, switching audio playback between the three frontal loudspeakers was not complex enough to convey a scenic impression or feeling of "togetherness", yet the relative simplicity of a three-loudspeaker layout per se might not necessarily result in lower experienced presence [11]. Other factors that could prove relevant for the internal formation of presence are the availability of prior knowledge about (changes in) talker locations [32] as well as visual cues of the test layout [15]. Lack of this prior knowledge in the present experiment might have confused participants, who probably expected a more static auditory scene due to the visible fixed loudspeakers placed in front of them.

The spatial mode also did not alleviate experienced difficulty in talker identification. The reason for this might be that only in half of the trials additional talker location cues were available to improve talker identification. Besides, the identification task was on average rated as "easy", even under degraded conditions; the behavioral performance increment

might therefore have been too small to be noticeable by participants and significantly affect their subjective judgments (ceiling effect). Future tests might consider increasing the complexity of the listening situation (e.g., number of talkers and voice similarity [22], [24]–[26]) and spatial reproduction mode (e.g., by use of Higher Order Ambisonics [12]) to check at what point spatialization affects subjective measures.

In a follow-up analysis, behavioral responses gathered online during the talker identification task will be examined. The analysis of behavioral response times might prove to be more sensitive in detecting effects of *reproduction mode* as well as potential interactions with *transmission quality*. Through this, variation in speed of talker identification at short time ranges (milliseconds), depending on *loudspeaker position* in the test layout, could eventually be uncovered.

VIII. Conclusion

A QoE model of loudspeaker-based speech reproduction was tested in a human talker identification task. Effects on subjective experience were revealed for *transmission quality* (clean, noisy, filtered)—including immersive aspects like voice naturalness, spatial presence and social presence—but not for *reproduction mode* (non-spatial, spatial). It is concluded that the employed listening scenario with two talkers and the spatial reproduction mode had probably not been complex enough for subjective benefits of spatialization to emerge. Nonetheless, a more subtle influence of spatial auditory cues on talker identification could not be precluded and might be detectable through future behavioral response time analysis.

Acknowledgment

This work was supported by the strategic partnership program between Technische Universität Berlin, Germany, and the Norwegian University of Science and Technology in Trondheim, Norway. The authors thank Tim Cato Netland for his technical support during the audio setup.

References

[1] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*, rev. ed ed. Cambridge, Mass: MIT Press, 1997.

[2] A. S. Bregman, *Auditory scene analysis: the perceptual organization of sound*. Cambridge, Mass: MIT Press, 1990.

[3] W. Zhang, P. Samarasinghe, H. Chen, and T. Abhayapala, "Surround by Sound: A Review of Spatial Audio Recording and Reproduction," *Applied Sciences*, vol. 7, no. 5, p. 532, 2017.

[4] S. Agrawal, A. Simon, S. Bech, K. Bærentsen, and S. Forchhammer, "Defining Immersion: Literature Review and Implications for Research on Immersive Audiovisual Experiences," in *Audio Engineering Society Convention 147*, 2019.

[5] J. Skowronek, K. Schoenenberg, and G. Berndtsson, "Multimedia Conferencing and Telemeetings," in *Quality of Experience*, S. Möller and A. Raake, Eds. Cham: Springer International Publishing, 2014, pp. 213–228.

[6] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A Spatial Audio Quality Inventory (SAQI)," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 984–994, 2014.

[7] C. S. Oh, J. N. Bailenson, and G. F. Welch, "A Systematic Review of Social Presence: Definition, Antecedents, and Implications," *Frontiers in Robotics and AI*, vol. 5, p. 114, 2018.

[8] P. Le Callet, S. Möller, and A. Perkis (Eds.), "Qualinet White Paper on Definitions of Quality of Experience," European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Version 1.2, March 2013.

[9] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*, ser. Signals and Communication Technology. Berlin, Germany: Springer, 2005.

[10] A. Silzle, "Generation of Quality Taxonomies for Auditory Virtual Environments by Means of a Systematic Expert Survey," *19th International Congress on Acoustics - ICA 2007 Madrid*, pp. 1–6, 2007.

[11] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, no. 9, pp. 651–666, 2002.

[12] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State," *Proc. of the IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013.

[13] M. Frank, F. Zotter, H. Wierstorf, and S. Spors, "Spatial Audio Rendering," in *Quality of Experience*, S. Möller and A. Raake, Eds. Cham: Springer International Publishing, 2014, pp. 247–260.

[14] R. Nicol, L. Gros, C. Colomes, M. Noisternig, O. Warusfel, H. Bahu, and B. F. G. Katz, "A Roadmap for Assessing the Quality of Experience of 3D Audio Binaural Rendering," in *EAA Joint Symposium on Auralization and Ambisonics*, Berlin, Germany, 2014.

[15] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. Lisbon, Portugal: IEEE, 2016, pp. 1–6.

[16] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acustica united with Acustica*, vol. 98, no. 5, pp. 804–810, 2012.

[17] A. Raake and J. Blauert, "Comprehensive modeling of the formation process of sound-quality," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. Klagenfurt am Wörthersee, Austria: IEEE, 2013, pp. 76–81.

[18] S. Uhrig, G. Mittag, S. Möller, and J.-N. Voigt-Antons, "Neural correlates of speech quality dimensions analyzed using electroencephalography (EEG)," *Journal of Neural Engineering*, vol. 16, no. 3, p. 036009, 2019.

[19] M. Schoeffler, A. Silzle, and J. Herre, "Evaluation of Spatial/3d Audio: Basic Audio Quality Versus Quality of Experience," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 75–88, 2017.

[20] M. Wältermann, A. Raake, and S. Möller, "Quality Dimensions of Narrowband and Wideband Speech Transmission," *Acta Acustica united with Acustica*, vol. 96, no. 6, pp. 1090–1103, 2010.

[21] J. J. Baldis, "Effects of spatial audio on memory, comprehension, and preference during desktop conferences," in *Proc. of the SIGCHI conference on Human factors in computing systems - CHI '01*. Seattle, Washington, United States: ACM Press, 2001, pp. 166–173.

[22] R. Kilgore, M. Chignell, and P. Smith, "Spatialized audioconferencing: What are the benefits?" in *Proc. of the 2003 Conference of the Centre for Advanced Studies on Collaborative Research*, 2003, pp. 135–144.

[23] A. Raake and C. Schlegel, "Auditory assessment of conversational speech quality of traditional and spatialized teleconferences," in *ITG Conference on Voice Communication [8. ITG-Fachtagung]*. VDE Verlag GmbH, 2008, pp. 1–4.

[24] A. Raake, C. Schlegel, K. Hoeldtke, M. Geier, and J. Ahrens, "Listening and Conversational Quality of Spatial Audio Conferencing," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, 2010.

[25] J. Skowronek and A. Raake, "Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing," in *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISCA, 2011, pp. 829–832.

[26] ——, "Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls," *Speech Communication*, vol. 66, pp. 154–175, 2015.

[27] A. Raake, *Speech Quality of VoIP: Assessment and Prediction*. Chichester, UK: John Wiley & Sons, Ltd, 2006.

[28] B. C. J. Moore and C.-T. Tan, "Perceived naturalness of spectrally distorted speech and music," *The Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 408–419, 2003.

[29] F. Köster, F. Schiffner, D. Guse, J. Ahrens, J. Skowronek, and S. Möller, "Towards a MATLAB Toolbox for Imposing Speech Signal Impairments Following the P.TCA Schema," in *Audio Engineering Society Convention 139*, 2015.

[30] ITU-T Recommendation P.851, *Subjective quality evaluation of telephone services based on spoken dialogue systems*. Geneva, Switzerland: International Telecommunication Union, 2003.

[31] S. Werner, F. Klein, and K. Brandenburg, "Influence of spatial complexity and room acoustic disparity on perception of quality features using a binaural synthesis system," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. Pylos-Nestoras: IEEE, 2015, pp. 1–6.

[32] D. S. Brungart and B. D. Simpson, "Cocktail party listening in a dynamic multitalker environment," *Perception & Psychophysics*, vol. 69, no. 1, pp. 79–91, 2007.