MEASURING USER ACCEPTANCE OF PROACTIVELY PLAYED TOURISTIC TEXTS IN AN IN-CAR VOICE ASSISTANT

Niklas Berensmeyer[1], *Stefan Hillmann*[1], *Wolfgang Maier*[2]

[1]*Technische Universität Berlin,* [2]*Mercedes-Benz AG*
berensmeyer@campus.tu-berlin.de

Abstract: User acceptance of proactive voice assistants remains insufficiently understood. We examine how language and content design influence user acceptance of TravelCompanion, a proactive in-car voice assistant providing touristic narrations. Using a mixed-method approach, we combined a 30-day automotive OEM field study with a controlled lab study to test four versions of TravelCompanion (neutral baseline, simpler sentence structures, higher informational density, friendlier tone). While the field study results showed only trends favoring simpler sentence structures, the lab study revealed significant preferences for both simpler sentence structures and friendlier tone. In contrast, higher informational density did not improve user acceptance. These findings show that specific language and content design characteristics significantly influence user acceptance of proactive in-car voice assistants.

1 Introduction

1.1 Background and Motivation

Voice assistants have evolved from simple command executors to sophisticated conversational partners capable of proactive interaction [1]. However, little is known about how users perceive and accept proactive behavior, particularly in automotive contexts [2]. This study addresses this gap by examining how different language and content design characteristics influence user acceptance of a real-world proactive in-car voice assistant domain that provides tourism-related information (TravelCompanion).

1.2 TravelCompanion

TravelCompanion is a fictional name for a real-world proactive voice assistant (VA) deployed on production level at a major automotive original equipment manufacturer (OEM). It provides drivers with tourism-related information about nearby points of interest (POIs) by proactively playing 20 to 60 second narrations (readouts) as they pass by.

In brief, POIs are retrieved together with additional contextual information from freely available online sources and subsequently clustered spatially. For each cluster, a readout text is generated using customizable prompts via a large language model. The generated text is then converted into speech using a state-of-the-art text-to-speech provider, and the resulting audio files are uploaded to cloud storage (off-board). When a TravelCompanion-enabled vehicle enters a cluster, the corresponding readout audio file is automatically downloaded and played in the vehicle (on-board).

2 Related Work

2.1 Proactive Voice Assistants

A VA is a software agent that facilitates human-computer interaction through a voice user interface [3]. VAs rely on spoken dialog systems to process and generate natural language [4]. The majority of current VAs operate on a reactive model, responding only to direct user input. While suitable for some use cases, this model has several limitations. Consequently, there is a growing shift toward proactive interaction, where VAs initiate dialog by themselves based on the environmental context [5]. This is often valuable in hands-free contexts and when cognitive load is already elevated, such as while driving [1]. However, designing such systems is challenging; if the timing or frequency of interventions is poorly calibrated, the system may become intrusive [6]. Yet, in automotive contexts, prior work indicates that proactive behavior is perceived as comparably positive to non-proactive interaction, particularly for driving-related tasks, while inducing lower increases in cognitive load than reactive approaches [1, 7].

2.2 Quality and Usability Evaluation for Spoken Dialog Systems in Vehicles

Evaluation of the quality and usability of SDSs typically relies on standardized questionnaires, interactive prototypes, and Wizard-of-Oz setups [8].

Quantitative measures frequently use Likert-scale-based instruments, such as user experience questionnaires [9]. Additional approaches include composite scores that capture dimensions such as information quality, semantic intelligence, and user satisfaction [10]. Although these methods enable efficient comparison, they necessarily abstract the multidimensional nature of user experience, often requiring complementary evaluation techniques [11]. Qualitative feedback collection is commonly conducted using open-ended questionnaire items, think-aloud protocols, semi-structured interviews, and diary studies. These methods provide deeper insight into user perceptions, expectations, and interaction strategies [12]. In practice, quantitative and qualitative data collection is often intertwined, leading to hybrid evaluation approaches, such as usability tests that integrate numerical metrics with subjective user feedback [13].

In-vehicle testing plays a central role in the development of new automotive features. It is often performed through field operational tests that evaluate system behavior under real-world driving conditions [14]. Various experimental designs can be applied and combined in this type of testing. Commonly used are A/B testing, which compares alternative feature versions to assess performance and user interaction, and continuous experimentation, where features are iteratively refined based on real-world usage data [15]. Driving simulators can serve as a complementary and, in some cases, equivalent evaluation environment, enabling controlled experimentation while yielding results comparable to field testing [16].

3 Method

3.1 Hypotheses

Considering the state of the art, we formulate three hypotheses on how language and content design influence user acceptance of TravelCompanion:

1. User acceptance of the TravelCompanion domain increases with a lower sentence structure complexity in a readout.

2. User acceptance of the TravelCompanion domain does not increase with a higher amount of details provided per POI in a readout.

Table 1 – Measured values for characteristics of each generated TravelCcompanion version. Bold values indicate the lowest value for LIX_{mean} and the highest values for W_{mean} and F_{mean}.

Version	LIX_{mean}	W_{mean}	F_{mean}
BASELINE	42.83	37.16	0.38
HYP1	**29.26**	35.86	0.31
HYP2	44.44	**78.38**	0.68
HYP3	36.51	53.24	**1.36**

3. User acceptance of the TravelCompanion domain increases with the use of a friendlier tone in a readout.

The first hypothesis builds on findings that simpler linguistic structures, as opposed to more complex ones, can enhance user acceptance in reactive interaction models [17, 18] and examines whether this also applies to proactive TravelCompanion readouts. It primarily concerns language design. The second hypothesis draws on research indicating that large amounts of information can overwhelm users, especially in driving scenarios [7, 19]. Unlike the first hypothesis, this hypothesis investigates content design. The third hypothesis inspects whether findings regarding the influence of emotional factors on the perception of reactive interactions [20, 21] are applicable to proactive TravelCompanion readouts. It relates to both language and content design.

3.2 Study Design

We apply a mixed-method research design, combining a real-world field study with a controlled lab study. In the field study, four systematically varied versions of TravelCompanion (neutral baseline, simpler sentence structures, higher informational density, friendlier tone) are deployed to a test vehicle fleet of an automotive OEM for 30 days to analyze natural user behavior. Objective technical usage data are logged and evaluated using quantitative metrics (abort rate and playback duration before abort). In the lab study, a driving simulator is used to collect subjective data with standardized questionnaires. Each driver passes through a cluster twice, hearing a different randomly chosen version of the readout each time. After each passage, the driver completes a questionnaire including the UEQ-S [22]. After both passages are completed, the driver additionally indicates a preference between the two versions.

3.3 TravelCompanion Version Characteristics

For the user studies, we generate four distinct versions of TravelCompanion: a neutral baseline version and one version that exhibits the content and language design characteristics of each hypothesis formulated. The baseline version is labeled BASELINE, while the labels HYP1 through HYP3 correspond to the respective version associated with each hypothesis. The readouts in each HYP version differ notably from BASELINE in one specific characteristic.

Table 1 shows the characteristics of each version. In line with the hypotheses, these differences concern either the complexity of sentence structure, quantified using the LIX readability index [23], the level of detail provided per POI, operationalized as the mean number of words used per POI, or the friendliness of the tone used, assessed via expert ratings on a five-point Likert scale ranging from -2, very unfriendly, to 2, very friendly. In all other aspects, the versions remain virtually consistent. This consistency is achieved by using nearly identical configurations in the generation process of each version, with the only variation being the prompts used to generate the readout texts. The underlying POI and cluster data remain identical across all versions.

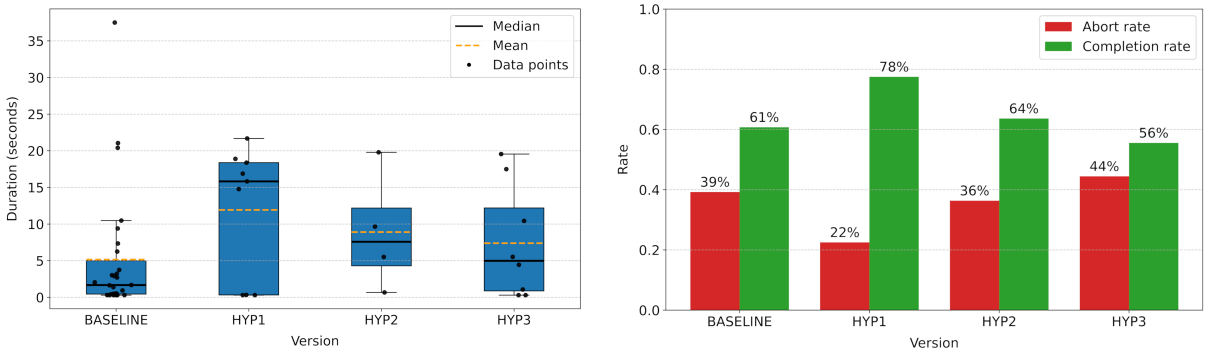

Figure 1 – Playback duration before abort, abort rate, and completion rate

3.4 Evaluation Metrics

To assess user acceptance per deployed TravelCompanion version, three evaluation metrics are based on the logged data of the field study: playback duration before abort, abort rate, and completion rate. The playback duration before abort metric reflects the duration a readout from a version was listened to before abortion by a driver. It is calculated using the start and abort timestamps provided in the data logged for each aborted readout. The abort rate indicates the proportion of readouts that were aborted by a driver out of all readouts played for a version. It is defined as the ratio of aborted readouts to the total number of readouts played in a version. Its complement is the completion rate, which resembles the percentage of readouts of a version that were played until the end.

The lab study questionnaire yields UEQ-S scores and version preference. The UEQ-S scores aim to evaluate different aspects of user acceptance of a version, and version preference aims to measure which of the compared versions are preferred by drivers.

4 Results

4.1 Field Study

The number of readouts per version and the corresponding distinct clusters are summarized in Table 2. As domain usage varied across test vehicles, readouts are not uniformly distributed between versions. Information on driver identities and passenger counts are not available. However, as the test fleet was accessible only to engineers and managers within the automotive OEM's research and development division, participants can be assumed to have had a general understanding of automotive technology, though not necessarily of the TravelCompanion domain.

Table 2 – Distribution of processable readouts and distinct clusters per version in the field study.

Version	Readouts played	Percentage (%)	Distinct clusters
BASELINE	79	53.4	45
HYP1	40	27.0	30
HYP2	11	7.4	6
HYP3	18	12.2	13

The left panel of Figure 1 shows the playback durations before abort, while the right panel shows abort and completion rates. Due to the limited sample size, no statistically significant effects of the version characteristics were observed on any evaluation metric. However, trends indicated that the use of simpler sentence structures reduced aborts.

4.2 Lab Study

A total of 58 simulated drives were conducted, with each participant completing one simulated drive. All participants were speech technology domain experts of the automotive OEM, familiar with the TravelCcompanion domain. Table 3 shows how often readouts of each version were played and how often versions were compared against each other.

Table 3 – Distribution of readout audios played and pairwise comparisons per version in the lab study. A total of 58 simulated drives were conducted, leading to 116 individual evaluations of versions collected.

Version	Readouts played	vs. BASELINE	vs. HYP1	vs. HYP2	vs. HYP3
BASELINE	29	–	11	9	9
HYP1	35	11	–	9	15
HYP2	23	9	9	–	5
HYP3	29	9	15	5	–

4.2.1 UEQ-S Scores

Figure 2 shows the UEQ-S scores recorded for each version. HYP3 achieved the highest scores in most dimensions, particularly in the hedonic dimensions, while HYP1 performed best in the remaining, primarily pragmatic dimensions. HYP2 consistently received the lowest scores among the HYP versions and was the only variant to record negative values in some dimensions, closely clustering with BASELINE.

Normality was violated for most UEQ-S dimensions, with only 6 of 32 version-dimension combinations meeting the Shapiro–Wilk criterion ($p \geq .050$). Accordingly, Kruskal–Wallis tests revealed a significant effect of version across all eight dimensions ($p < .050$), with medium to large effect sizes, strongest for confusing–clear and inefficient–efficient. Post-hoc Dunn tests showed that HYP3 differed significantly from BASELINE in seven dimensions, including three hedonic ones, while HYP1 differed from BASELINE in three pragmatic dimensions, all with medium to large effects. In contrast, effects involving HYP2 were generally small or negligible.

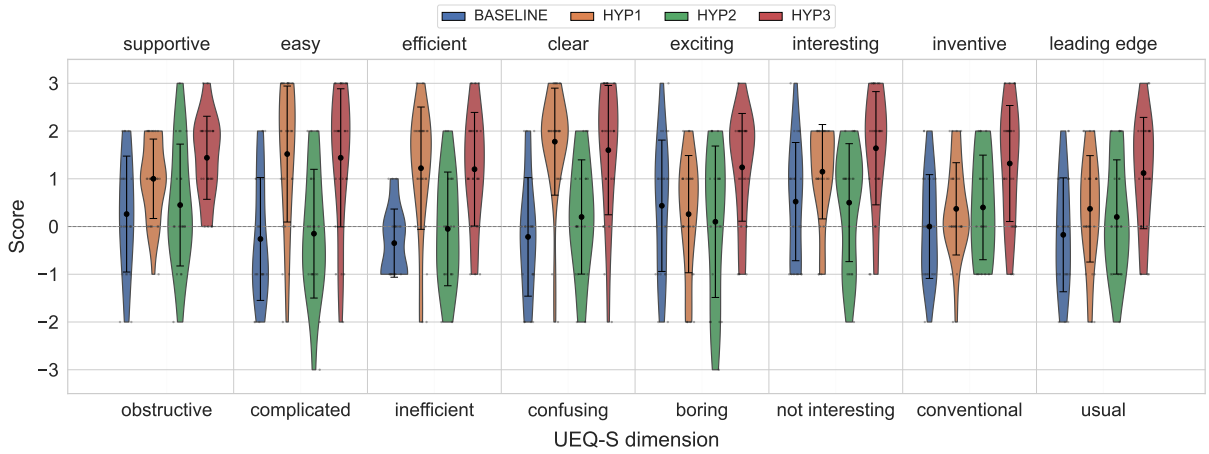

Figure 2 – UEQ-S scores

4.2.2 Version Preference

Figure 3 presents the version preference results as a heatmap, showing for each ordered version pair the percentage of comparisons in which the row version was preferred over the column version. Undecided responses are included. Thus, opposing preferences do not necessarily sum up to 100%. HYP3 was consistently preferred over all other versions, with participants up to seven times more likely to favor it over BASELINE. An even stronger preference over BASELINE was observed for HYP1, which was chosen ten times more often. While both HYP3 and HYP1 achieved high preference rates, HYP3 showed a slight advantage in their direct comparison. BASELINE was never preferred over any other version, but performed comparatively better against HYP2 than against the remaining HYP versions.

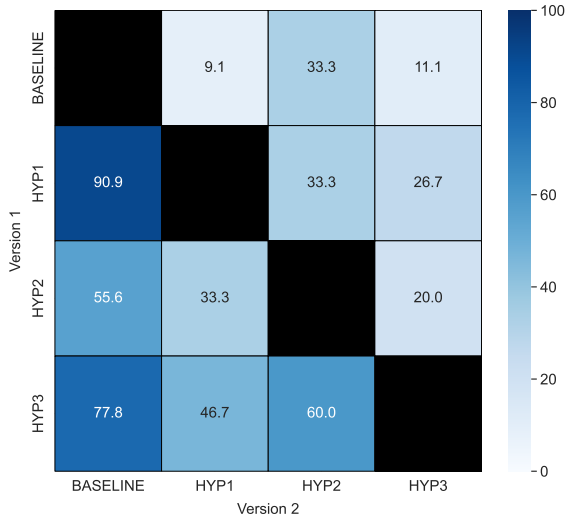

Figure 3 – Version preference heatmap

5 Discussion

In the field study, HYP1 showed markedly longer playback durations before abort and a substantially lower abort rate compared to BASELINE, suggesting higher user tolerance for longer interactions when sentence structure complexity is reduced. However, these effects were not statistically significant. In the lab study, HYP1 was rated significantly clearer, easier, and more efficient than BASELINE, with large effect sizes. It was also consistently preferred over BASELINE in direct comparisons. Overall, the lab study provides strong evidence supporting Hypothesis 1, aligning with prior work linking lower syntactic complexity to reduced cognitive load and higher acceptance in driving contexts [7, 19].

HYP2 achieved slightly longer playback durations before abort than BASELINE in the field study, but abort and completion rates were comparable and non-significant. In the lab study, HYP2 did not differ significantly from BASELINE across all UEQ-S dimensions. It was also not clearly preferred in direct comparisons. These results support Hypothesis 2, suggesting that increasing the amount of information per POI does not enhance user acceptance, consistent with prior findings that excessive information can overwhelm users in driving scenarios [7, 19].

While the field study results for HYP3 were inconclusive, the lab study revealed clear effects. HYP3 significantly outperformed BASELINE across all UEQ-S dimensions, with medium to large effect sizes. It was also strongly preferred over any version in direct comparisons. These findings provide robust support for Hypothesis 3 and are consistent with previ-

ous work demonstrating that a friendly conversational style increases user acceptance of voice assistants [20, 21].

6 Conclusion

We aimed to examine the impact of systematically varied language and content design characteristics on user acceptance of proactively played touristic texts in an in-car voice assistant context. The results show that specific language and content design characteristics can significantly influence user acceptance of the TravelCompanion domain and likely other proactive touristic in-car voice assistants. They suggest that the use of simpler sentence structures and a friendlier tone increases user acceptance, whereas no clear positive effect can be demonstrated for the use of higher informational density.

References

[1] DU, H., X. FENG, J. MA, M. WANG, S. TAO, Y. ZHONG, Y.-F. LI, and H. WANG: *Towards proactive interactions for in-vehicle conversational assistants utilizing large language models*. In *Proc. IJCAI-24*, pp. 7850–7858. 2024. doi:10.24963/ijcai.2024/869.

[2] NOTHDURFT, F., S. ULTES, and W. MINKER: *Finding appropriate interaction strategies for proactive dialogue systems: An open quest*. In *Proc. MMSYM 2015*, pp. 73–80. 2015. URL `https://ep.liu.se/en/conference-article.aspx?article_no=10&issue=110`.

[3] HOY, M. B.: *Alexa, siri, cortana, and more: An introduction to voice assistants*. *Medical Reference Services Quarterly*, 37(1), pp. 81–88, 2018. doi:10.1080/02763869.2018.1404391.

[4] YANG, H., A. STOLCKE, and L. P. HECK: *Spoken conversational agents with large language models*. In *Proc. EMNLP 2025: Tutorial Abstracts*, pp. 7–8. ACL, Suzhou, China, 2025. doi:10.18653/v1/2025.emnlp-tutorials.3.

[5] BÉRUBÉ, C., M. NISSEN, R. VINAY, A. GEIGER, T. BUDIG, A. BHANDARI, C. R. PE BENITO, N. IBARCENA, O. PISTOLESE, P. LI, A. BIN SAWAD, E. FLEISCH, C. STETTLER, B. HEMSLEY, S. BERKOVSKY, T. KOWATSCH, and A. B. KOCABALLI: *Proactive behavior in voice assistants: A systematic review and conceptual model*. *Computers in Human Behavior Reports*, 14, p. 100411, 2024. doi:10.1016/j.chbr.2024.100411.

[6] MIKŠÍK, O., I. MUNASINGHE, J. ASENSIO-CUBERO, S. REDDY BETHI, S.-T. HUANG, S. ZYLFO, X. LIU, T. NICA, A. MITROCSAK, S. MEZZA, R. BEARD, R. SHI, R. W. M. NG, P. A. M. MEDIANO, Z. FOUNTAS, S.-H. LEE, J. MEDVESEK, H. ZHUANG, Y. ROGERS, and P. SWIETOJANSKI: *Building proactive voice assistants: When and how (not) to interact*. *arXiv*, abs/2005.01322, 2020. doi:10.48550/arXiv.2005.01322.

[7] SCHMIDT, M., W. MINKER, and S. WERNER: *User acceptance of proactive voice assistant behavior*. In *Proc. ESSV 2020*, pp. 18–25. Springer, 2020. URL `https://www.essv.de/pdf/2020_18_25.pdf`.

[8] CLARK, L., P. DOYLE, D. GARAIALDE, E. GILMARTIN, S. SCHLÖGL, J. EDLUND, M. AYLETT, J. CABRAL, C. MUNTEANU, and B. R. COWAN: *The state of speech in hci: Trends, themes and challenges*. *Interacting with Computers*, 31(4), pp. 349–371, 2019. doi:10.1093/iwc/iwz016.

[9] LAUGWITZ, B., M. SCHREPP, and T. HELD: *Construction and evaluation of a user experience questionnaire*. In A. HOLZINGER (ed.), *Proc. USAB 2008*, vol. 5298 of *LNCS*, pp. 63–76. Springer, 2008. doi:10.1007/978-3-540-89350-9_6.

[10] ZWAKMAN, D. S., D. PAL, T. TRIYASON, and C. ARPNIKANONDT: *Voice usability scale: Measuring the user experience with voice assistants*. In *Proc. iSES 2020*, pp. 308–311. IEEE, 2020. doi:10.1109/iSES50453.2020.00074.

[11] LINDGAARD, G. and C. DUDEK: *What is this evasive beast we call user satisfaction? Interacting with Computers*, 15(3), pp. 429–452, 2003. doi:10.1016/S0953-5438(02)00063-2.

[12] LAZAR, J., J. H. FENG, and H. HOCHHEISER: *Research Methods in Human–Computer Interaction*. Morgan Kaufmann, Cambridge, MA, 2 edn., 2017.

[13] KHAYYATKHOSHNEVIS, P., S. TILLBERG, E. LATIMER, T. AUBRY, A. FISHER, and V. MAGO: *Comparison of moderated and unmoderated remote usability sessions for web-based simulation software: A randomized controlled trial*. In M. KUROSU (ed.), *Human-Computer Interaction. Theoretical Approaches and Design Methods*, Lecture Notes in Computer Science, pp. 232–251. Springer, 2022. doi:10.1007/978-3-031-05311-5_16.

[14] BARNARD, Y., S. INNAMAA, S. KOSKINEN, H. GELLERMAN, E. SVANBERG, and H. CHEN: *Methodology for field operational tests of automated vehicles*. *Transportation Research Procedia*, 14, pp. 2188–2196, 2016. doi:10.1016/j.trpro.2016.05.234.

[15] ROS, R. and P. RUNESON: *Continuous experimentation and a/b testing: A mapping study*. In *Proc. 4th RCoSE*, pp. 35–41. ACM, 2018. doi:10.1145/3194760.3194766.

[16] WANG, Y., B. MEHLER, B. REIMER, V. LAMMERS, L. A. D'AMBROSIO, and J. F. COUGHLIN: *The validity of driving simulation for assessing differences between in-vehicle informational interfaces: A comparison with field testing*. *Ergonomics*, 53(3), pp. 404–420, 2010. doi:10.1080/00140130903464358.

[17] LUGER, E. and A. SELLEN: *Like having a really bad pa: The gulf between user expectation and experience of conversational agents*. In *Proc. CHI 2016*, pp. 5286–5297. ACM, 2016. doi:10.1145/2858036.2858288.

[18] HWANG, A., N. OZA, C. CALLISON-BURCH, and A. HEAD: *Rewriting the script: Adapting text instructions for voice interaction*. *arXiv preprint*, 2023. doi:10.48550/arXiv.2306.09992.

[19] MECK, A.-M., C. DRAXLER, and T. VOGT: *How may i interrupt? linguistic-driven design guidelines for proactive in-car voice assistants*. *International Journal of Human-Computer Interaction*, 40(22), pp. 7517–7531, 2023. doi:10.1080/10447318.2023.2266251.

[20] NICULESCU, A., B. VAN DIJK, A. NIJHOLT, H. LI, and S. L. SEE: *Making social robots more attractive: The effects of voice pitch, humor and empathy*. *Int. J. of Social Robotics*, 5(2), pp. 171–191, 2013. doi:10.1007/s12369-012-0171-x.

[21] SNYDER, E. C., S. MENDU, S. S. SUNDAR, and S. ABDULLAH: *Busting the one-voice-fits-all myth: Effects of similarity and customization of voice-assistant personality*. *International Journal of Human-Computer Studies*, 180, p. 103126, 2023. doi:10.1016/j.ijhcs.2023.103126.

[22] SCHREPP, M., A. HINDERKS, and J. THOMASCHEWSKI: *Design and evaluation of a short version of the user experience questionnaire (ueq-s)*. *Int. J. of Interactive Multimedia and AI*, 4(6), 2017. doi:10.9781/ijimai.2017.09.001.

[23] BJÖRNSSON, C. H.: *Läsbarhet*. No. 6 in Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. Liber, Stockholm, Sweden, 1968.