ERRONEOUS REACTIONS OF VOICE ASSISTANTS "IN THE WILD" – FIRST ANALYSES

Lea Kisser[1]*, Ingo Siegert*[1]

[1] *Institute of Information and Communication Engineering, Otto von Guericke University Magdeburg, Germany*
lea.kisser@st.ovgu.de, ingo.siegert@ovgu.de

Abstract: Voice assistants are increasingly dominating everyday life and represent an easy way to perform various tasks with minimal effort. The areas of application for voice assistants are diverse and range from answering simple information questions to processing complex topics and controlling various tasks. However, current voice assistants very quickly reach their limits in everyday operation, as individual differences in their use, especially in the formulation of requests, encounter a uniform way of providing information. This is especially evident when voice assistants respond differently than expected by the user. These "erroneous reactions" are responded to with individually different strategies by the users regarding the content, but also lexical and prosodic variations.
In the present contribution, we rely for the first time on a large data collection (approx. 40h) in which a large number of users interact with a voice assistant in a non-laboratory setting. Specifically, we focus on interactions with incorrect responses from the system. Thereby Incorrect responses are understood to be system messages that contain an error message or that failed to appear at all.

1 Introduction

Voice assistants are increasingly dominating everyday life and represent an easy way to perform various tasks with minimal effort. This development has led to a rapidly growing user base for commercial voice assistants [1]. This development is not limited to the US market, as another study showed that in 2019 60% of all Germans have already used voice assistants [2].

The application for voice assistants reaches through different areas, such as Smart Home Control, Mobile Assistance, or Operating Systems. Hereby users actually employ their voice assistants mostly for voice search, setting up an alarm clock and reminders, controlling other devices, or listening to music [3]. One reason for the popularity of voice assistants is their given naturalness of speaking as a form of communication. However, current voice assistants very quickly reach their limits in everyday operation, as individual differences in their use, especially in the formulation of requests, encounter a uniform way of providing information [4]. This is especially evident when voice assistants respond differently than expected by the user. These "erroneous reactions" are responded to with individually different strategies by the users regarding the content, but also lexical and prosodic variations.

In the present contribution, we rely for the first time on a large data collection (approx. 40h) in which a large number of users interact with a voice assistant unconstrained in a non-laboratory setting. Specifically, we focus on interactions with incorrect responses from the system. Thereby Incorrect responses are understood to be system messages that contain an error message or that appeared not to correspond to the user's intention.

2 Related Work

Current research on voice-assistants is mainly focused on three aspects: (1) technological improvements providing better voice recognition/understanding, human-like speech output, or in future adding affective computing abilities, cf. [5], (2) privacy and security issues to increase the trust in and prevent abuse of voice assistants, cf. [6, 7], (3) conversational analyses to investigate and explain the usage of these devices [8].

This contribution focuses on the third block of aspects and concentrates on struggles in the interaction with modern voice assistants, especially on erroneous reactions from the technical system. Only a few studies already investigated the interaction with (modern) voice assistants and analyzed the struggles within the dialogs. Hereby, studies directly focusing on erroneous conversations used a WoZed setup with just a few participants in a lab environment. The authors of [8] investigated conversational repair using a manipulated Alexa device to "make a mistake" in understanding the participant and afterwards let the assistant perform a correction. Another study conducted between-subject user experiments with 30 older adults using voice assistants designed with different error handling strategies [9]. Hereby, again a WoZ-setup was used. Both studies could show that self-repair and the identification of errors in the interaction is a crucial part of the interaction and that multiple error handling strategies are beneficial. A further recent paper conducted usability tests to investigate users' error handling behavior for real voice assistant responses for 20 Brazilian users [10]. The authors identified 8 different voice assistant's response types, of were 7 are considered erroneous as well as 6 different user strategies to cope with the errors. Among them, a change in the prosody, a variation on the amount of interaction, and repeating a command were mostly observed.

Other studies using a commercial voice assistant mostly focus on the question of how users change their way of interaction, regarding accommodation [11] or addressee behavior [12].

Unconstrained, non-lab interactions are analyzed so far only for a very limited user base, of either private households, see [13] or [14], or public student interactions, see [15]. The first two studies examine the use of modern voice interfaces (Amazon Echo) in the everyday life in five house-holds and nine households respectively with about a month of recording. Besides having unconstrained conversations and showing that these devices are embedded into the life of the participants, the studies analyzed erroneous responses rather anecdotic and from a user's perspective on communication satisfaction. The latter study investigated how students interact with Amazon Alexa in a public academic space. Lopatovska & Oropeza then alalysed the 79 recorded sessions in terms of user satisfaction and usability at all. They observed that Alexa sometimes responds to participants' questions with irrelevant information. But, in the majority of erroneous interactions, Alexa admitted its inability to help or remained silent.

One major drawback of all of these studies is that either the task was pre-defined and the participants tried to solve the task with the given means, even if that include coping with imperfect systems or the number of participants and setting was too limited to draw a general conclusion. Furthermore, most of these studies focused on the participants' perspective instead of a systematic analysis of the voice assistant's responses.

3 Dataset

As dataset for the current study, the Voice Assistant Conversations in the wild (VACW) dataset has been used [16]. This dataset consists of unconstrained and unscripted interactions of users and a modern commercial voice assistant. The recordings took place during the science exhibition of the "MS Wissenschaft". The "MS Wissenschaft" is a yearly touring exhibition with a distinct topic. In 2019 it has visited 31 cities in Germany and Austria. Thus a lot of different regional dialects are present.

The exhibit was developed to show the lacking functionality of today's intelligent voice

assistants. It was inspired by the quiz "Who Wants to Be a Millionaire". Visitors had to answer random questions presented on a screen together with possible answers and could make use of an Amazon Alexa to answer the questions, which were designed in such a way that Alexa is not able to answer directly. Thus, the visitor has to ask for partial steps or has to reformulate the question. The exhibit did not have supervision or further instruction given to the visitors. Thus due to the setup, the visitors were free in their decision how to interact with the voice assistant, how to react to its responses, which question to ask, and when to abort or stop the interaction [16].

Simultaneously, this exhibit was used to record unconstrained interaction data. It could be assumed that due to the exhibition and its gamification character the visitors' shyness to interact with a voice assistant was reduced and therefore many different visitors interacted with the voice assistant. In addition, the exhibition explicitly invited visitors to try and play around with the exhibits. For the data recording, an unmodified version of the Amazon Echo Input was used and the data was extracted from the specifically created Amazon user account for the exhibit.

During the 126 days of the exhibition, a total of about 37k utterances with a total duration of about 40h was collected. For each speech utterance the timestamp of the interaction and the transcription of the user query (the way it was transcribed/understood by Alexa) as well as Alexa's response, if applicable, was stored, as well. The recording had the full approval of the data security officer of the University.

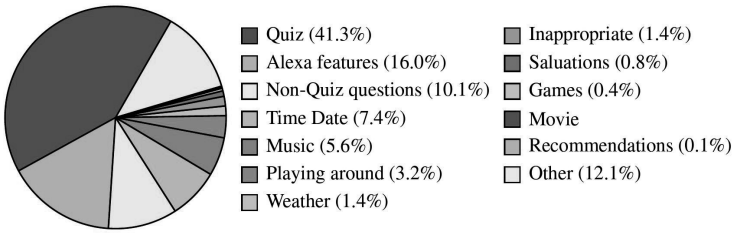

Figure 1 – Pie diagram of the types of visitor interactions present in the VACW dataset, 100% corresponds to 29k utterances.

To give an overview of which topics or functions are addressed by the visitors, a simple regex search was used, see Figure 1. It can be concluded that aside from the quiz-related questions, similar topics as in [15] are present. Requests for Alexa features and time/date-related requests are occurring quite often. The topics salutations, games, movie/TV, and recommendations are uttered comparable seldom in this dataset. What is striking in VACW, is the large proportion of inappropriate requests (swear words, insults), which even includes a few racist expressions.

4 Methods

After manually cleaning the VACW dataset of incomplete entries and queries not intended for the voice assistant, approximately 29,000 entries were left for further analyses. First, entries comprising incorrect responses were identified. This includes approx. 46% of all entries. For the remaining entries, the assistant's responses were categorized according to the type of feedback.

Subsequently, based on the auditory impression (prosody or acoustic conspicuousness) and the transcripts, possible reasons for the incorrect reaction were sought on the one hand, and users' strategies to overcome the error and their success was analyzed.

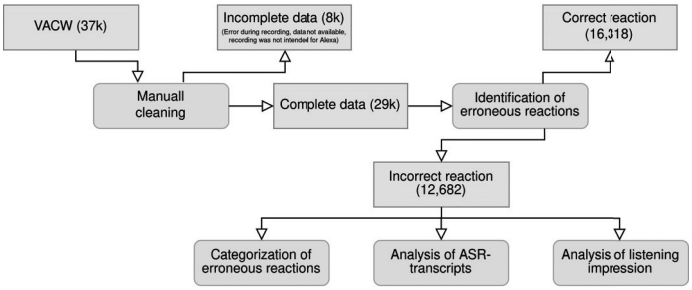

Figure 2 – Experimental procedure to analyse the erroneous interactions. Numbers in brackets denotes the amount of utterances.

5 Results

5.1 Categorization of error responses

Three types of false responses could be identified in the present material. For 34% of all user queries, there was no system response [1], for 6.96% of all entries, there was a standardized error response, in the form of pre-programmed response without any real reference to the situation/user input, and for 4.88% of all entries, there was a specific error response, where the system response refers to the previous user input. Furthermore, for 2.72% of all requests, some kind of confirmation requests were given by the voice assistant. Thus, only half of all requests (51.42%) could be directly answered by the voice assistant after the initial request. For an overview, see the left pie diagram in Figure 3.

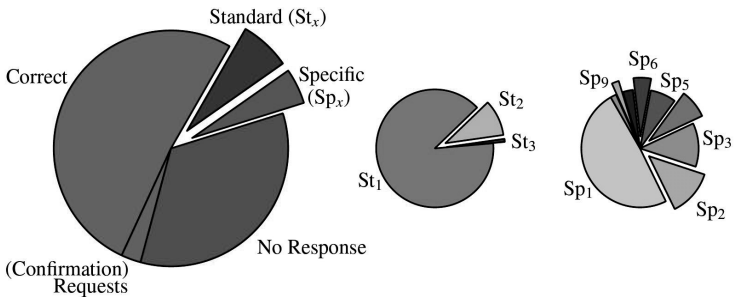

Figure 3 – Distribution of error responses in total (left pie diagram), standardized error messages (middle pie diagram) and specific error messages (right pie diagram).

Regarding the **standardized error messages**, they can be roughly divided into three aspects, see the middle pie diagram in Figure 3. For the biggest group of all standardized error messages, it can be assumed that this is related to errors in the NLU and the correct recognition of the users' intent ($\mathbf{St}_1$), e.g. *"Das weiß ich leider nicht"*, *"Ich bin mir leider nicht sicher"*, or *"Ich weiß nicht, wie ich dir dabei helfen kann"*. The next group of standardized errors messages comprise problems in (acoustically) understanding the user ($\mathbf{St}_2$) (9.9% of all standardized error messages), e.g. *"Entschuldigung, es sieht so aus, als hätte ich dich nicht richtig gehört. Bitte versuche es erneut"* or *"Entschuldigung, das habe ich nicht verstanden. Bitte versuche es erneut"*. The smallest group (0.8% of all standardized error messages) comprise errors in fetching the information ($\mathbf{St}_3$) either due to connection problems or commands out of the scope of Alexa, e.g. *"Entschuldigung, etwas ist schiefgelaufen"*, *"Entschuldigung, da ist etwas schiefgelaufen. Bitte wiederhole die Frage"*, or *"Tut mir leid, aber das ist nicht möglich"*.

[1] It should be noted here that all queries not intended for Alexa were already sorted out in the previous step.

The group of **specific error messages** refers to reactions for which a command could not be executed, but the intention behind the command was recognized (even if only supposedly). For this category of errors individualized error responses, mostly containing explanations of the specific problem, are given for different categories. The distribution of related topics is heavily influenced by the visitors and (maybe) by the type of interaction and thus not representative for other studies. The topics, in descending order, are: $\mathbf{Sp}_1$ media (48.94%), $\mathbf{Sp}_2$ skills (12.69%), $\mathbf{Sp}_3$ smart home (11.99%), $\mathbf{Sp}_4$ quiz-related questions (7.97%), $\mathbf{Sp}_5$ location (7.12%), $\mathbf{Sp}_6$ Alexa account (4.87%), $\mathbf{Sp}_7$ Chefkoch skill (2.82%), $\mathbf{Sp}_8$ language (1.98%), $\mathbf{Sp}_9$ shopping (1.62%).

Further **specific non-erroneous responses** that stood out included expressions of politeness. These include words like *"please"* and *"thank you"* or *"excuse me"*, *"sorry"* and *"unfortunately"* and could be observed for 9.17% of all answers. Furthermore, it was observed that certain responses were issued by Alexa in particular when it was confronted with inappropriate or insulting statements by the user. If the user called the voice assistant stupid, the feedback function in the Alexa app was mostly referred to. At other points Alexa responded to insults with *"Das ist nicht nett von dir"* or *"Entschuldigung. Das war nicht meine Absicht"* These responses only occurred for common insults with expressions such as *"stupid"*, *"dumb"*, or *"ugly"*. Alexa did not respond to insults that included phrases such as *"whore"*, *"fuck you"*, *"bitch"*, or indicated that Alexa did not know how to respond. It could be assumed that these insults were not recognized as such by the system. Moreover, the actual response strategy does neither support solving a miscommunication nor does it help to appease the user.

5.2 Analysis of visitors' ASR-Transcripts

The user requests are available only as automatic ASR-transcripts from the Alexa system and due to the large number of utterances in the database, a complete manual correction phase has not been conducted. Randomly selected sub-samples of this dataset have been used in [17]. In this context, the recognition accuracy of other ASR APIs was tested based on manual corrections. Transferred to the entirety of all entries, it can be assumed that the WER is between 17 and 25%. Afterwards, the transcripts of the user request leading to an erroneous reaction are manually screened for peculiarities. It must be noted that these analyses leave a lot of room for interpretation, as the exact intention of the users remains unclear since they interacted with Alexa mindlessly and without the possibility of follow-up questioning. Particularly noticeable findings will be briefly summarized in the following. This is not a complete list and can be regarded as a working hypothesis for further analyses.

- Frequent repetitions of the same or similar questions indicate that the user's intention was not recognized/answered.
- In addition, an interruption of the interaction was often observed, although this cannot necessarily be attributed to frustration but also to the character of the exhibition to try out the exhibits. This is an interesting aspect for follow-up studies to analyze when and how a conversation was ended/broken off.
- It also proved problematic when users restructured the sentence during their inquiry. Alexa was then no longer able to capture the original statement of the sentence.
- Even when users wanted to cancel their request by interrupting their utterance and using words like *"Stopp"* or *"Abbrechen"*, this could not always be detected by the system. Instead, Alexa tried to interpret the incomplete command, which accordingly resulted in incorrect responses.
- As previously mentioned, some users use offensive language. However, this is only recognized as such by Alexa for a few expressions and Alexa reacts with an apology or an indication of inappropriate behavior. But for many expressions, the insulting character is not recognized. These cases have so far been rarely reported in previous studies and lend themselves to further research on expressions of frustration in dialogue.

- Alexa requires certain parameters (slots) to execute certain commands. When they are missing and Alexa asks for them, it expects the information for this parameter in the upcoming answer. This led to inconsistencies in communication at some points, as users sometimes gave answers that did not follow this pattern. If, for example, instead of responding to Alexa's inquiry the topic of conversation was changed, this could not be detected. Alexa's NLU continued to analyze the answer to match the missing parameters, and thus missed the changed context of the new query.
- Furthermore, it could be observed that the users had problems leaving an unintentionally activated skill again afterwards.

5.3 Analysis of Listening Impressions

In addition to the text-based analyses, voice recordings were manually inspected for further anomalies. In order to analyze the different solution strategies, error messages having one or more follow-up user requests were examined. For this purpose, all related interactions with the same speaker and similar language content were identified manually. This resulted in 804 dialogs, with the number of consecutive utterances ranging from 2 to 14.

Particular attention was paid to phonostylistic and prosodic speech characteristics. The former include speech tempo, intonation, sound assimilation, speech stress, and accentuation [18]. Within prosody, particular attention was paid to changes in volume, pitch, and pausing [19]. The observations reported in the following are purely subjective.

Phonostylistic highlights: Users frequently spoke more clearly after an error message. In particular, they switch from dialect to standard German. The use of dialect seems to be one of the main reasons why Alexa misunderstood words, besides words that are basically difficult to understand. Further, the queries of children, who are disproportionately represented in this data set, were also more likely to be misunderstood than those of adults. Especially, users tried to solve misunderstandings by speaking slower than before. A clear pronunciation seems to be successful only to a limited extent. Especially, unknown/complicated expressions leading to a wrong recognition, pose a general comprehension problem for the speech assistant.

Prosodic highlights: The analysis of the fundamental frequency of the voice recordings showed that users tend to change their pitch in order to manage conflict situations with the voice assistant. In particular, conversations that included three or more consecutive changes in the fundamental frequency could be observed just for a minority of these cases. Furthermore, only for a few of them (15%), a correct response was received, afterwards. For the remaining conversations, either a termination of the interaction or another solution strategy (mostly reformulation) was observed. Thus, another means has predominantly resorted in order to escape the problem.

Furthermore, it could be observed that the perceived volume varies strongly between successive recordings, but partly also within the recordings. On the one hand, the perception suggests that this is due to a different distance to the microphone. In other cases, it sounds more like the pronunciation itself is getting louder, or a kind of speech enhancement has been activated.

Another observation, for many quiz questions users tend to read out the question first and then verbally list the answer choices. The recordings now show that users made a pause for breath between the end of the question and the beginning of the enumeration. This pause was interpreted by Alexa as the end of the query so that the given answers were no longer taken into account and the answers were therefore incorrect and thus led to misunderstandings.

Other factors: By analyzing the recordings, it was further noticed that especially during group interactions and by misunderstandings of the activation signaling misinterpretations during the interaction arise. In group interactions, for example, it was often observed that users interrupted each other or talked in a mixed-up manner. The speech recognition could no longer reliably differentiate between background noise/background conversations and requests. The acoustic

signal, which indicates that Alexa is actively listening, also frequently led to disruptions in the interaction. On the one hand, users waited for this signal after saying the activation word before starting with the actual command. If, in this case, no utterance follows the activation, Alexa ends the interaction again, even though that is not the user's intention. It is conceivable that the users simply could not perceive the signal due to the loud background noise during the exhibition. On the other hand, users were occasionally disturbed in their query when the acoustic signal occurred while they were already beginning to formulate their query. The resulting restructuring of the query or confusion during the request causes many errors.

6 Conclusion and Outlook

This paper presents a first in-depth analysis of unconstrained interactions with modern voice assistants. In these first analyses, the focus was on erroneous responses of the system and their systematic analysis. The goal was to establish hypotheses for further investigations. With regard to the user strategies in reaction to an erroneous response, it can be stated that changes in prosody usually show no success. From this, it can be concluded that the communication of errors needs improvement in many cases. It seems that users lack an idea of how speech assistants must be used and what the reasons are for an incorrect response. However, changes in prosody can perhaps be used to automatically detect faulty system responses.

Regarding the transcripts, it can be stated that in general, the ASR performance is quite good. It was further noticed that frustration is often manifested by offensive expressions and that a change in the sentence structure is not successful if the ASR engine is responsible for the error, the users are not aware of this and are also not informed. What was striking in comparison to the analyses of the recordings is that especially the background noise and the presence of several speakers at once seem to promote an erroneous response. This so far has not been present in other studies conducted in a usually more silent lab environment with just one speaker. Furthermore, the acoustic signal from the voice assistant often seems to cause irritation, which then led to errors in the formulation. The listening impression showed that users use various strategies to react to an incorrect response. Frequently, the pitch is varied, the speaking tempo is reduced, the volume is increased, and a more "dialect-free"/clearer pronunciation is used. However, it turns out that for the analyzed interactions these strategies have only limited success. In most cases, prosodic variations did not lead to a successful request. Uncertainties during sentence construction seem to promote an erroneous response.

In general, it seems as if the communication of errors has so far only been solved in a very basic manner and that the users receive concrete feedback only in very few cases. Furthermore, it seemed as if standard system responses and reactions are randomly selected from a set of pre-defined responses hindering a proper understanding. Together with a sometimes missing visualization of the interaction, the freedom that a speech-based interaction actually offers is limited to inflexible and rigid command input. Thus, it is still crucial for a successful interaction that the user gives clear instructions and plans in advance on how to formulate a request.

References

[1] KINSELLA, B.: *Nearly 90 million u.s. adults have smart speakers, adoption now exceeds one-third of consumers*. voicebot.ai, 2020. URL `https://perma.cc/336P-2C77`. [Online; posted 28-Apr-2020].

[2] SPLENDID RESEARCH GMBH: *Studie: Digitale sprachassistenten und smart speaker*. Januar 2019. URL `https://www.splendid-research.com/de/studie-digitale-sprachassistenten.html`. [Online; posted 2021].

[3] SERPIL TAS, R. A.: *Nutzung von sprachassistenten in deutschland.* In *Sprachassistenten - Anwendungen, Implikationen, Entwicklungen : ITG-Workshop : Magdeburg*, p. 39. 2020.

[4] VALLI, A.: *Notes on natural interaction.* Tech. Rep., University of Florence, Italy, 2007.

[5] PADMANABHAN, J. and M. J. J. PREMKUMAR: *Machine learning in automatic speech recognition: A survey. IETE Technical Review*, 32(4), pp. 240–251, 2015. doi:10.1080/02564602.2015.1010611.

[6] DUBOIS, D. J., R. KOLCUN, A. M. MANDALARI, M. T. PARACHA, D. CHOFFNES, and H. HADDADI: *When Speakers Are All Ears: Characterizing Misactivations of IoT Smart Speakers.* In *Proc. of the Privacy Enhancing Technologies Symposium (PETS)*. 2020.

[7] SIEGERT, I.: *Speaker anonymization solution for public voice-assistant interactions – presentation of a work in progress development.* In *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, pp. 80–82. 2021.

[8] CUADRA, A., S. LI, H. LEE, J. CHO, and W. JU: *My bad! repairing intelligent voice assistant errors improves interaction.* In *Proc. ACM Hum.-Comput. Interact.*, pp. 1–24. 2021. doi:10.1145/3449101.

[9] LIN, W., H.-C. CHEN, and H.-P. YUEH: *Using different error handling strategies to facilitate older users' interaction with chatbots in learning information and communication technologies. Frontiers in Psychology*, 12, 2021. doi:10.3389/fpsyg.2021.785815.

[10] MOTTA, I. and M. QUARESMA: *Users' error recovery strategies in the interaction with voice assistants (vas).* In *Proc. of the 21st Congress of the International Ergonomics Association (IEA 2021)*, pp. 658–666. Springer International Publishing, Cham, 2022.

[11] RAVEH, E., I. SIEGERT, E. STEINER, I. GESSINGER, and B. MÖBIUS: *Three's a crowd? effects of a second human on vocal accommodation with a voice assistant.* In *INTERSPEECH 2019*, pp. 4005–4009. 2019.

[12] SIEGERT, I. and J. KRÜGER: *"Speech Melody and Speech Content Didn't Fit Together"—Differences in Speech Behavior for Device Directed and Human Directed Interactions*, pp. 65–95. Springer International Publishing, Cham, 2021. doi:10.1007/978-3-030-51870-7_4.

[13] PORCHERON, M., J. E. FISCHER, S. REEVES, and S. SHARPLES: *Voice Interfaces in Everyday Life*, p. 1–12. New York, NY, USA, 2018. doi:10.1145/3173574.3174214.

[14] MAVRINA, L., J. SZCZUKA, C. STRATHMANN, L. M. BOHNENKAMP, N. KRÄMER, and S. KOPP: *"alexa, you're really stupid": A longitudinal field study on communication breakdowns between family members and a voice assistant. Frontiers in Computer Science*, 4, 2022. doi:10.3389/fcomp.2022.791704.

[15] LOPATOVSKA, I. and H. OROPEZA: *User interactions with "Alexa" in public academic space. Proc. of the Association for Information Science and Technology*, 55, pp. 309–318, 2018. doi:10.1002/pra2.2018.14505501034.

[16] SIEGERT, I.: *"Alexa in the wild" – Collecting Unconstrained Conversations with a Modern Voice Assistant in a Public Environment.* In *Proc. of the 12th LREC*, pp. 608–612. ELRA, Marseille, France, 2020. URL `https://www.aclweb.org/anthology/2020.lrec-1.76`.

[17] SIEGERT, I., Y. SINHA, O. JOKISCH, and A. WENDEMUTH: *Recognition Performance of Selected Speech Recognition APIs – A Longitudinal Study*, pp. 520–529. Springer, Cham, 2020.

[18] HIRSCHFELD, U., B. NEUBER, and E. STOCK: *Was ist eine gute Aussprache?* Dudenverlag, Mannheim, Leipzig, Wien, Zürich, 2007.

[19] NEUBER, B.: *Leistungen der Suprasegmentalia fuer das Verstehen, Behalten und die Bedeutungs(re)konstruktion.* PETER LANG Frankfurt am Main, 2002.