Effects of Prosodic Variations on Accidental Triggers of a commercial Voice Assistant

Ingo Siegert

Mobile Dialog Systems, Institute for Information Technology and Communications,
Otto von Guericke University Magdeburg, Germany

`ingo.siegert@ovgu.de`

Abstract

The use of modern voice assistants has rapidly grown and they can be found in more and more households. By design, these systems have to scan every sound in their surroundings waiting for their respective wake-word before being able to react to the users' commands. The drawback of this method is that phonetic similar expressions can activate the voice assistant and thus speech utterances or whole private conversations will be recorded and streamed to the cloud back-end for further processing. Many news articles and scientific work reported on inaccurate wake-word detection. Resulting in at least a user's confusion or at worst security breaches. The current paper is based on a broader analysis of phonetic similar accidental triggers conducted by Schönherr et al., they presented a systematic analysis to detect accidental triggers, using a pronouncing dictionary and a weighted, phone-based Levenshtein distance. In this work, the previously identified accidental triggers are recorded by several speakers under various conditions to investigate the influence of phonetic variances (i.e. intonation and speaking/articulation rate) on the robustness of accidental triggers in a real-world environment.

Index Terms: voice assistants, accidental triggers, computational paralinguistics

1. Introduction

In recent years, the market for commercial voice assistants has been continuously rising. Resulting in a nearly doubled user-base in the last three years among the U.S. adults [1, 2] and an increasing number of owners using voice assistants in their daily routine [3]. This is mostly because of their simplicity and pervasiveness in activation and operation. But, this simplicity has implications on how these (commercial) devices operate: To properly react to an intended activation, voice assistants continuously analyze every sound in their surroundings to detect a so-called wake-word. Only then, the voice assistant is activated and ready to process the user's request by sending the recorded utterance to a cloud infrastructure, responsible for the transcription and interpretation of the request [4, 5]. This solution is mainly used due to limitations in the on-device computing power (embedded devices prohibit a full analysis of the audio stream on the device itself) and privacy concerns (cloud processing is only conducted when the wake-word intended a user request). Thus, the proper detection of a wake-word is a crucial point in the whole processing pipeline. Unfortunately, the development of a wake-word detection with low false acceptance rates to avoid accidental activations and low false rejection rates to avoid users constantly repeating the wake-word in order to activate the voice assistant is a quite challenging task [6, 7, 8, 9]. Some researchers already reported that by accidental triggers voice assistants can be 'hacked'. Vaidya et al. [10] demonstrated that the

misinterpretation of "cocaine noodles" as "OK Google" enables unauthorized commands (sending a text or dialing a number). Similar approaches of using imperfect transcriptions to invoke malicious skills having similar-sounding names are presented in [11, 12]. This issue is not only fatal due to the adverse usage of the voice assistant, but also, due to privacy reasons even when the accidental activation is emerging from the users itself: the recording and sending of a private conversation of a family, for which Amazon later justified that the voice assistant reacted due to a word in the background conversation sounding like the wake-word [13]. Already the streaming of the recorded data of an accidentally activated voice assistant is mostly seen as a privacy threat [14, 15, 16], as employees of the manufacturers are listening to voice recordings in order to transcribe them and improve the voice assistants' performance [17, 18].

To systematically analyze the phenomenon of accidental triggers, the authors of [19] implemented an approach to automatically craft accidental trigger candidates and evaluated the resistance of various commercial voice assistants to such accidental triggers. In their research, they used a text-to-speech (TTS) service to validate their identified accidental trigger candidates and published a dataset of 132 identified English accidental triggers for 4 different voice assistants to foster future research on this topic. As more than 60% of the identified accidental triggers are related to the wake-words of Amazon's ALEXA[1], this paper limited the analyses to the 89 accidental trigger candidates identified for Amazon's ALEXA. Based on this, the current paper deals with the following research questions: 1) How do real recordings of the previously identified accidental trigger candidates by [19] behave? 2) Can the accidental activation be traced back to specific phonetic variations, e.g. F0 differences?

Thereby this paper contributes the following novel aspects to the research community: Investigations on real recordings of accidental triggers for voice assistants and giving phonetic insights into the accidental activation of voice assistants.

2. Experimental Setup

2.1. Utilized Accidental Triggers

The accidental triggers utilized in this study were originally identified in [19]. The authors presented a method to fully automatically craft accidental trigger candidates. Their approach is based on the assumption that possible candidates should have a similar pronunciation (expressed by similar phones) to the wake-words. Utilizing a weighted phone-based Levenshtein distance [19, 20] to identify potential candidates from the Fisher corpus [21] version of the Carnegie Mellon University pronouncing dictio-

---

[1]This can be partly attributed to the fact that Amazon offers 4 wake-words and partly to the fact that Google, Apple, and Microsoft use a two-word activation phrase.

nary [22], together with a cross-validation using TTS-generated trigger candidates with cross-checking the occurrences in real-world conversation recordings using the CHiME dataset [19, 23], the authors of [19] identified 89 accidental trigger candidates for the four wake-words available for Amazon's voice assistants. This approach guarantees that the identified accidental trigger candidates are occurring within English conversations. The trigger candidates are constructed of different 1-, 2-, and 3-gram word sequences, see Table 1 and Fig. 2 for some examples.

Table 1: *Number of accidental triggers for Amazon's ALEXA identified by [19, see Table VII for a full list], separated by 1-, 2- or, 3-gram word sequences.*

Wake Word	1-gram	2-gram	3-gram
Alexa	7	10	5
Computer	16	12	10
Echo	1	8	3
Amazon	2	11	4

2.2. Conducted Online Data Collection

To increase the phonetic variability of the accidental trigger candidates, the current study recorded human speech samples rather than TTS-voices. Due to the pandemic limitations and with the aim to collect a broad variety of speakers, an online data collection using the SoSci Survey was utilized [24]. This allows participants to contribute to the data collection remotely, but also increases the risk of poor quality recordings, see Sections 3 and 4.1.1. After being informed about the purpose of this data collection and her/his consent to store and process the recordings, each participant had the task to record the previously identified 89 accidental trigger candidates. Furthermore, the participants are asked to indicate their mother language, age, sex, and recording device. The whole survey took about 15-20 Minutes.

2.3. Accidental Trigger Measurement Setup

The measurement setup is similar to the setup used in [19] and consist of the smart speaker (ALEXA Echo Dot 3rd Generation), a loudspeaker with nearly linear frequency response (distance to smart speaker approx 1 m), and a webcam to monitor the activity of the smart speaker (using the brightness and color information), a computer playing back the recorded speech samples, evaluating the activation-monitor, and providing a local WiFi network for the smart speaker. Each sample is played twice and any smart speaker LED activity is logged. The log file includes the filename of the activated speech sample and the timestamp of activation.

3. Collected Accidental Trigger Recordings

The SoSci survey was available from 19.09.2020 until 31.01.2021 and in total 251 users took part. After a manual inspection of the recordings (a substantial number of samples recorded for a participant, i.e. more than 2/3 of all samples) 76 participants uttering 5 251 accidental trigger samples remain.

The recordings comprise 33 male and 43 female speakers with a mean age of 26.2 years (std. deviation of 8.24 years, minimum 17 years, maximum 57 years). In total 13 different mother languages (including variants) are indicated, with a majority of Russian (RUS), Czech (CZE), English (ENG), German (GER), Indo-Aryan (INC), and Chinese (CHI), see Fig. 1.

According to the different recording devices, most participants used their laptop microphone (54.5%), followed by smart-
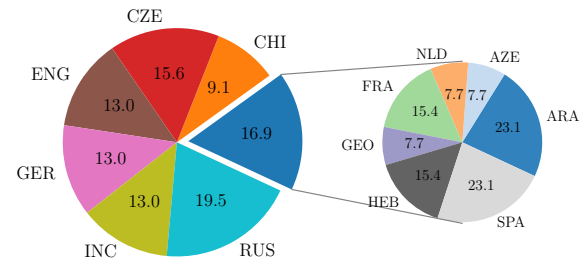

Figure 1: *Participants' mother language distribution, according to ISO 639-2 codes.*

phone or consumer headsets (both 15.6%). Only a few participants used a professional headset or a tablet microphone, 5.2% and 3.9%, respectively. Four participants (5.2%) have not specified their recording device.

4. Measured Accidental Trigger Results

The accidental triggers are identified using the measurement setup described in Section 2.3. In total, only 150 samples (2.87%) accidentally activated the voice assistant. These samples comprise 35 (46.1%) different trigger candidates. Scaled to the number of existing voice assistants and conducted interactions, this is a factor that should not be underestimated regarding the real number of accidental activations [1, 2]. An additionally conducted loudness normalization (cf. [25]) of the recorded samples only shows very small differences in the accidental activation. Therefore, the analyses for the unnormalized samples are reported here.

The distribution of observed activations for the investigated accidental trigger candidates is depicted in Fig. 2. It can be seen that in general 1-gram trigger candidates (i.e. less complex) have a higher activation rate. But also some 2-gram candidates lead to a remarkable percentage of false activations. In comparison, 3-gram candidates tend to be more robust.

Regarding the associated speakers, a high variation can be observed. From the total of 76 speakers only for 22 (28.9%), the recorded trigger samples do not lead to any false activation. The remaining 54 speakers (71.1%) fall apart into two groups: a) false activations below average (1-2 accidental triggers per participant) and b) false activations above average (3-8 accidental triggers per participant). Thus, it can be assumed that either the trigger candidates pose different robustness against phonetic variations or that other (external) factors are the cause of this phenomenon. Possible external factors are speaker characteristics and recording quality. Hence, additional in-depth analyses of the recorded trigger samples are needed.

4.1. In-Depth analyses of Accidental Trigger Recordings

In the following, it is examined to which extent various factors differ in the three groups: a) no activation (noA), b) below average, i.e. low activation (loA), and above average, i.e. high activation (hiA).

4.1.1. Recording Quality

At first, it is aimed to exclude that differences in the recording quality influence the accidental activation. Therefore, the reported recording devices the calculated Signal-to-Noise ratio (SNR) as well as the loudness values will be compared between the three groups (noA, loA, and hiA).
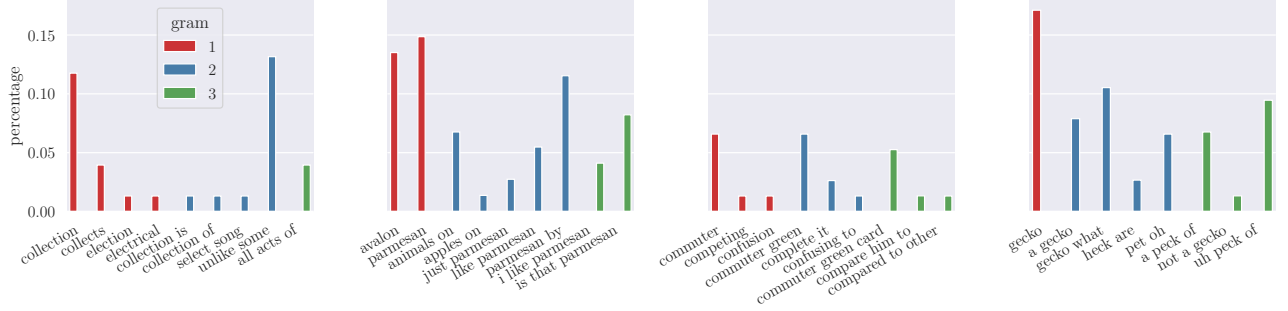
Figure 2: *All accidental triggers caused an activation in our measurement setup, ordered by n-gram size and trigger command (left to right: "Alexa, "Amazon", "Computer" and "Echo"). For the remaining 41 trigger candidates no activation could be observed.*

Regarding the reported recording device, nearly no differences could be observed between the three groups. Around 55% of the participants in each group used a laptop microphone. Both, smartphone microphones and consumer headsets were used by roughly 1/6 of all participants in each group. A tablet microphone was used by approx 4% of the participants for each group. The only (slight) difference could be observed for participants using a professional headset, only occurring in the group with no activation (4.5% one participant) and high activation (12%, 3 participants). But these numbers are too low to draw reliable conclusions.
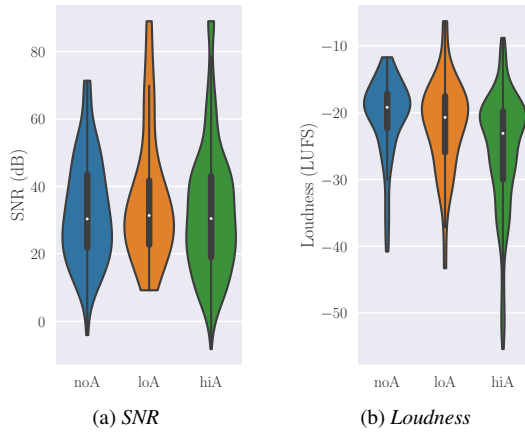

(a) *SNR*   (b) *Loudness*

Figure 3: *SNR and Loudness values regarding the three groups of no (noA), low (loA), and high (hiA) accidental activation.*

SNR and loudness are important indices of the recording quality [26, 27]. To calculate the SNR silent and non-silent parts of each recorded sample were automatically identified based on a forced alignment using the MAUS tool [28]. The loudness values were extracted using the tool *mp3gain*, which provides Loudness Units relative to Full Scale (LUFS) values [25]. The distribution of SNR and loudness of the recorded samples regarding the three groups of accidental activation is depicted in Fig. 3. As it could be assumed from the similar distribution of recording devices in the three groups, the SNR also reveals no substantial differences between the groups. It furthermore has to be noted that the SNR within a speaker varies by approx. 10dB. But apart from that, no substantial differences could be observed.

4.1.2. Speaker Characteristics

Speaker characteristics in this investigation comprise age, sex, and mother language. The details are given in Table 2. Regarding

age, no substantial differences can be found between the three groups, the mean age is 24.7 years, 25.3 years, and 28.9 years for noA, loA, and hiA, respectively. The sex is nearly equally distributed in the noA (11m and 12f) and hiA (12m and 13f). Only for the loA group, the sex is highly imbalanced with 10 male and 19 female speakers. Nevertheless, it can be assumed that sex does not influence the accidental trigger robustness. The distribution of mother languages regarding the three speaker groups is quite diverse, but there is no trend regarding a specific language family identifiable that could distinguish accidental activations.

Table 2: *Speaker characteristics for the three groups of accidental trigger activations.*

	age [years] mean (std)	sex m/f	mother language 4 most frequent
noA	24.7 (4.32)	11/12	RUS, INC, GER, CZE
loA	25.3 (6.84)	10/19	CZE, RUS, ENG, INC
hiA	28.9 (11.50)	12/13	GER, RUS, CHI, ENG

As the technical inspection, as well as the speaker characteristics, do not reveal compelling reasons for higher accidental activation for some speakers, it can be assumed that there exist some prosodic/phonetic reasons. Therefore, in the following, it is analyzed whether there are differences in the pronunciation of specific triggers by speakers where the accidental trigger candidate leads to an activation compared to speakers where this is not the case. It has to be noted that due to the experimental design (participants record their voice under very different not fully documentable conditions), there is still room for factors that could not be analyzed, e.g. distance from the microphone, health condition, recording setup quality, degradations included due to compression.

4.2. Prosodic analyses of Selected Accidental Triggers

From previous research, it is known that speakers, when asked about it, most often state that they speak more slowly towards technical systems than with human interlocutors [29, 30] and change their speaking behavior, especially speech melody/intonation, when talking to machines [29, 31]. Thus, speaking slower and changing intonation are analyzed with the assumption that they are the most distinguishable characteristics when talking to machines and humans.

4.2.1. Speech/Articulation Rate

Therefore, the first distinguishable factor could be the "speed of speaking", which is usually measured either as speech rate or

articulation rate. Both are defined as the number of production units per unit time, where the speech rate includes and the articulation rate excludes pauses [32]. As in the present study, isolated words or groups of words were analyzed, the calculation for both measures is based on phonemes. The timings of the phonemes and the pauses for 2-gram and 3-gram accidental triggers were automatically identified based on a forced alignment using the MAUS tool [28]. The results are depicted in Fig. 4. Regarding the three groups of noA, loA, and hiA, no substantial differences in the resulting articulation rate or speech rate could be revealed. Thus it can be assumed that neither the speaking speed nor pauses between the words (for 2/3 gram triggers) influenced the observed accidental activation.
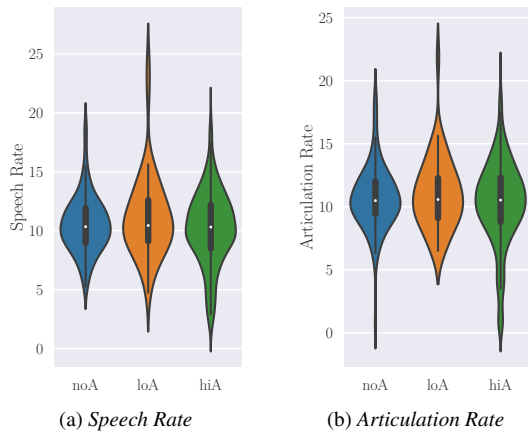

(a) *Speech Rate*   (b) *Articulation Rate*

Figure 4: *Speech Rate and Articulation Rate regarding the three groups of no (noA), low (loA), and high (hiA) accidental activation.*

4.2.2. Intonation

It can further be assumed that the differences in the accidental activation are resulting from differences in the intonation, which is usually measured as the contour of the fundamental frequency (F0). The F0 values are extracted for each phoneme of the investigated triggers using the OpenSmile toolkit [33].
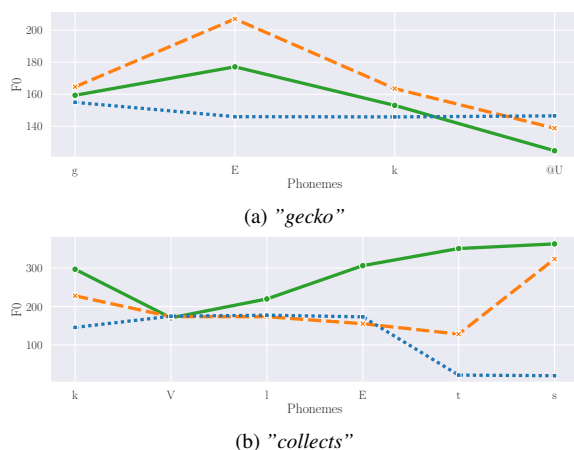

(a) *"gecko"*

(b) *"collects"*

Figure 5: *Differences between F0 contours, depicted for two accidental triggers, for* noA *,* loA *, and* hiA *.*

In contrast to the general assumption that device-directed speech is of low intonation (cf. [34]), the results, exemplarily

depicted in Fig. 5, reveal the contrary indication. For "gecko" the F0 contour has a much smaller range for trigger candidates not leading to an activation, for "collects" the F0 contour for the accidental trigger activations is directed upwards, while it is directed downwards (with a smaller range) for those that are not leading to an activation. This behavior can be found for nearly all trigger candidates in the three groups of noA, loA, and hiA. This is further supported by the calculated standard deviation, of 70.2 Hz, 119.5 Hz, and 100.1 Hz regarding the F0 range for noA, loA, and hiA, respectively.

5. Limitations

Due to the (non-standardized) recording setup, high variations are imposed into the data, but due to the Corona-Pandemic, this was the only option to collect this amount of speech data. The variance in the recordings does not allow the investigation of the influence of specific factors. But this allows that this wide spectrum of influences can be investigated at all and at least trends and hypotheses can be raised. Furthermore, only isolated speech has been investigated, the effect of continuous speech with slurring of words is left for future investigations.

Also, to allow reproducibility, varying rooms and acoustic environments were not investigated, although it has already been shown that they influence the (emotional) prosodic information [35]. Instead, it was focused on real recordings of accidental triggers in a comparable setup to [19]. Therefore, the findings are based on the US-English language, with pronunciations from mainly non-native speakers. Since it was not possible to get a deeper insight into the wake-word detection system, many of the results are based only on observations and indirect analyses derived from them. Access to the activation function of the detection system would be needed for deeper analyses.

6. Conclusions

This paper analyzed the intonation variations, in terms of "speed of speaking" and F0 values, of trigger candidates for commercial voice assistants, identified by [19] on the accidental activation. It, therefore, relies on real (remote) recordings from different speakers under different conditions.

Regarding the two research questions, it can be stated that real recordings depict different influences on accidental activation and that for our data the accidental activation can be traced back to distinct differences in the intonation. Other characteristics of the recordings, as recording quality, speaker characteristics, or "speaking speed" do not show substantial differences regarding an accidental activation. Especially the fact that a higher intonation variety leads to a higher accidental activation was somehow surprising, as one would assume the opposite behavior from the usual usage reports of voice assistants, cf. [29, 30].

Follow-up research should investigate the revealed connection between intonation and accidental activation in a more controlled setup, once this is possible again, as well as the investigation of accidental activations in continuous speech as this is left out in the current setup.

7. Acknowledgements

8. References

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

[2] Amazon, "Press release – customers shopped at record levels this holiday season with billions of items ordered worldwide – plus customers purchased tens of millions of amazon devices," posted 26-Dec-2019. [Online]. Available: https://perma.cc/7L3P-C86L

[3] S. Kleinberg, "5 ways voice assistance is shaping consumer behavior." think with Google, posted Jan-2018. [Online]. Available: https://perma.cc/U2Y2-Q4WN

[4] J. Konzelmann, "Chatting up your google assistant just got easier." The Keyword, blog.google, posted Jun-21-2018. [Online]. Available: https://perma.cc/CAT9-R4PR

[5] M. Wu, "Alexa scientists present two new techniques that improve wake word performance." Amazon Science, April 20186, [Online; posted 18-April-2018].

[6] O. Akhtiamov, I. Siegert, A. Karpov, and W. Minker, "Using complexity-identical human- and machine-directed utterances to investigate addressee detection for spoken dialogue systems," *Sensors*, vol. 20, no. 9, p. 2740, 2020.

[7] S. H. Mallidi, R. Maas, K. Goehner, A. Rastrow, S. Matsoukas, and B. Hoffmeister, "Device-directed utterance detection," in *Proc. of the INTERSPEECH'18*, Hyderabad, India, 2018, pp. 1225–1228.

[8] X. Tong, C.-W. Huang, S. H. Mallidi, S. Joseph, S. Pareek, C. Chandak, A. Rastrow, and R. Maas, "Streaming ResLSTM with causal mean aggregation for device-directed utterance detection," 2020.

[9] K. Gillespie, I. C. Konstantakopoulos, X. Guo, V. T. Vasudevan, and A. Sethy, "Improving device directedness classification of utterances with semantic lexical features," in *Proc. of the IEEE ICASSP-2020*, 2020, pp. 7859–7863.

[10] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: Exploiting the gap between human and machine speech recognition," in *9th USENIX Workshop on Offensive Technologies (WOOT 15)*, Washington D.C., USA, Aug. 2015.

[11] D. Kumar, R. Paccagnella, P. Murley, E. Hennenfent, J. Mason, A. Bates, and M. Bailey, "Skill squatting attacks on Amazon Alexa," in *27th USENIX Security Symposium (USENIX Security 18)*, Baltimore, USA, Aug. 2018, pp. 33–47.

[12] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian, "Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems," in *IEEE Symposium on Security and Privacy*, 2019, pp. 1381–1396.

[13] G. Horcher, "Woman says her amazon device recorded private conversation, sent it out to random contact." KIRO7, updated 25-May-2018. [Online]. Available: https://perma.cc/8CG9-3M3G

[14] H. Chung, M. Iorga, J. Voas, and S. Lee, "Alexa, can i trust you?" *Computer*, vol. 50, no. 09, pp. 100–104, sep 2017.

[15] D. J. Dubois, 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.

[16] N. Malkin, J. Deatrick, A. Tong, P. Wijesekera, S. Egelman, and D. Wagner, "Privacy attitudes of smart speaker users," *Privacy Enhancing Technologies*, vol. 2019, no. 04, pp. 250–271, 2019.

[17] C. Gartenberg, "Apple apologizes for Siri audio recordings, announces privacy changes going forward." The Verge, posted 31-Jul-2019. [Online]. Available: www.theverge.com/2019/8/28/20836760/

[18] C. Lecher, "Google will pause listening to EU voice recordings while regulators investigate." The Verge, posted 01-Aug-2019. [Online]. Available: www.theverge.com/2019/8/1/20750327/

[19] L. Schönherr, M. Golla, T. Eisenhofer, J. Wiele, D. Kolossa, and T. Holz, "Unacceptable, where is my privacy? exploring accidental triggers of smart speakers," *arXiv cs.CR 2008.00508*, 2020.

[20] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, no. 1, p. 31–88, Mar. 2001. [Online]. Available: https://doi.org/10.1145/375360.375365

[21] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Proc. of the 4th LREC'04*, Lisbon, Portugal, May 2004.

[22] K. A. Lenzo, "Carnegie Mellon Pronouncing Dictionary (CMUdict) - Version 0.7b,," Nov. 2014. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[23] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. Interspeech 2018*, 2018, pp. 1561–1565. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1768

[24] D. J. Leiner, "SoSci Survey (Version 3.1.06)," 2019, available at https://www.soscisurvey.de.

[25] EBU, "Loudness normalisation and permitted maximum level of audio signals," European Broadcasting Union, EBU Recommendation R128, 2020. [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I/en

[26] S. Das and P. Choudhury, "Evaluation of perceived speech quality for VoIP codecs under different loudness and background noise condition," in *Proc. 21st Int. ACM Conf. on Distributed Computing and Networking*, 2020.

[27] N. B. H. Croghan, K. H. Arehart, and J. M. Kates, "Quality and loudness judgments for music subjected to compression limiting," *The Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1177–1188, 2012.

[28] F. Schiel, "MAUS goes iterative," in *Proc. of the 4th LREC'04*, Lisbon, Portugal, May 2004.

[29] I. Siegert and J. Krüger, ""speech melody and speech content didn't fit together"—differences in speech behavior for device directed and human directed interactions," in *Advances in Data Science: Methodologies and Applications*, G. Phillips-Wren, A. Esposito, and L. C. Jain, Eds. Cham: Springer International Publishing, 2021, pp. 65–95.

[30] ——, "How do we speak with ALEXA - subjective and objective assessments of changes in speaking style between HC and HH conversations," *Kognitive Systeme*, vol. 1, p. s.p., 2018.

[31] V. Silber-Varod, A. Lerner, and O. Jokisch, "Prosodic plot of dialogues: A conceptual framework to trace speakers' role," in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds. Cham: Springer International Publishing, 2018, pp. 636–645.

[32] J. Koreman, "Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 582–596, 2006.

[33] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. of the ACM MM-2010*, 2010.

[34] T. Tsai, A. Stolcke, and M. Slaney, "Multimodal addressee detection in multiparty dialogue systems," in *Proc. of the IEEE ICASSP-2015*, April 2015, pp. 2314–2318.

[35] J. Höbel-Müller, I. Siegert, R. Heinemann, A. F. Requardt, M. Tornow, and A. Wendemuth, "Analysis of the influence of different room acoustics on acoustic emotion features," in *Elektronische Sprachsignalverarbeitung 2019. Tagungsband der 30. Konferenz*, Dresden, Germany, 2019, pp. 156–163.