The Legal and Ethical issues Workshop @LREC2022, pages 44–47
Marseille, 20 June 2022

Public Interactions with Voice Assistant – Discussion of Different One-Shot Solutions to Preserve Speaker Privacy

Ingo Siegert[1], Yamini Sinha[1], Gino Winkelmann[1], Oliver Jokisch[2], Andreas Wendemuth[3]
[1]Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany
[2]Cybersecurity and Data Management, HSF University, Meissen, Germany
[3]Cognitive Systems, Otto von Guericke University Magdeburg, Germany
siegert@ovgu.de, yamini.sinha@ovgu.de, gino.winkelmann@st.ovgu.de, oliver.jokisch@hsf.sachsen.de

Abstract

In recent years, the use of voice assistants has rapidly grown. Hereby, above all, the user's speech data is stored and processed on a cloud platform, being the decisive factor for a good performance in speech processing and understanding. Although usually, they can be found in private households, a lot of business cases are also employed using voice assistants for public places, be it as an information service, a tour guide, or a booking system. As long as the systems are used in private spaces, it could be argued that the usage is voluntary and that the user itself is responsible for what is processed by the voice assistant system. When leaving the private space, the voluntary use is not the case anymore, as users may be made aware that their voice is processed in the cloud and background voices can be unintendedly recorded and processed as well. Thus, the usage of voice assistants in public environments raises a lot of privacy concerns. In this contribution, we discuss possible anonymization solutions to hide the speakers' identity, thus allowing a safe cloud processing of speech data. Thereby, we promote the public use of voice assistants.
Keywords: voice assistant, public recordings, speaker anonymization

1. Introduction

The topic of data protection is becoming increasingly important today. In the field of voice assistants, there have been many discussions in recent years dealing with the protection of personal data (Schönherr et al., 2020; Siegert et al., 2021). Most current applications of voice assistants focus on the use of own assistant systems in private environments, but there are many applications possible where voice assistants can be employed in public spaces (i.e., museum guides, self-shopping support, etc (Porcheron et al., 2018; Lopatovska and Oropeza, 2018; Steven et al., 2017). In this context, the discussion is intensified on the one hand, as speech applications have rapidly grown and by their convenience of use along with outstanding speech understanding even in difficult acoustic environments. On the other hand, by the fact that for the first time, technical systems not just process personal data that can be used to identify users but directly use the voice as personal (biometric) data for the interaction. Furthermore, when using voice assistants, users are getting aware (often for the first time) that their data is processed in the cloud. The implications of a privacy breach in regard to speech data and possible solutions are extensively presented in (Tomashenko et al., 2021; Siegert et al., 2020a).

From a legal perspective, the protection of personal data is often limited to the country or jurisdiction in which the person lives. Other jurisdictions may have incompatible privacy policies which forbid to transfer any private data between them (Nautsch et al., 2019). In recent years, attempts have been made to draft agreements and decisions between the European Union and the United States of America in the field of data protection. The goal of each of these negotiations has been to reach an agreement on how to securely transfer personal data of citizens of European member states to the United States of America, based on the European Data Protection Directive. The resolutions negotiated to date have been progressively declared unworkable (European Court of Justice (Grand Chamber), 2020).

Especially, the aim of the GDPR to "enhance individuals' control and rights over their personal data and to simplify the regulatory environment" could cause practical implications on the use of voice assistants for public applications. Regarding the processing of the content of the transmitted data itself, this may not be crucial (at least) for information-providing systems, as users can be informed beforehand that their request will be processed and no personal data should be given. Recent field studies using a public voice assistant service showed that users do not tend to disclose private information (Siegert, 2020). But, in terms of the voice data itself, this is not possible, as users can be clearly identified by their voice, and a (recognizable) voice profile can be created. Thus, it could be possible to not only identify users based on their voice profile, but also to carry out additional voice analyses, such as the current mood or affect, which are shown to have an influence on the shopping experience.

A possible solution to prevent the identification of users by their voice is to rely on anonymization techniques. Anonymization has the advantage that truly anonymized data are not subject to the GDPR and as such can in principle be freely processed. But according to the GPDR, anonymization should take into account two parts, 1) that it is irreversible and 2) that it is done in such a way that it is impossible (or extremely impractical) to identify the data subject (WP29 (Article 29 Data Protection

Working Party), 2014). What does that imply for the use of voice assistants in public spaces? To answer that question, we formulated several assumptions:

- Users usually make only a few requests to the voice assistant.
- The anonymization should work regardless of the speaker's language, accent, sex, or gender.
- The anonymization should be fast, i.e., without a distracting delay in the interaction.
- The anonymization should run locally, independently of the voice assistant, and should also provide an audio stream.

In the following, three anonymization techniques are presented and their implication and outcomes are discussed. Hereby, we limit the utilized methods to work without a training phase, as we identified that as the most critical issue during the interaction with voice assistants. We believe that for a satisfying speech-based interaction, one-shot anonymization should be aimed so that the users can directly utter their commands which will be anonymized "on-the-fly".

2. Utilized Anonymization Techniques

2.1. McAdams coefficient

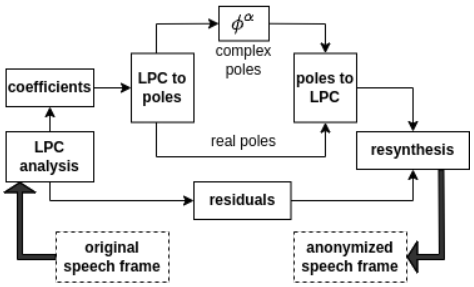

Figure 1: Pipeline of McAdams coefficient-based speaker anonymization, taken from (Patino et al., 2020). The angle ϕ of non-zero imaginary poles are raised to the power of McAdams coefficient α.

The anonymization technique using the McAdams coefficient (α) adjusts the timbre or spectral envelope, by frame-wise shifting the formant positions (Patino et al., 2020). Formats describe the spectral maximum resulting from an acoustic maximum of the human vocal tract and their location defines the peculiarity of specific phonemes (i.e. a unit of sound that distinguish words, as letters would do for written language). Due to anatomic differences in the vocal tract of humans, the formants for the same phoneme of different humans underlie specific variations. Therefore, slight variations to the formant positions obscure the speaker characteristics but preserve the produced sound, cf. (Siegert, 2015). Therefore, the formant positions have to be estimated from the original's speaker formants (i.e. speech analysis), then the formant shift has to be applied and afterwards, the speech including the shifted formants has to be re-produced (i.e. speech synthesis).

To get an estimation of the source (i.e. residuals) and filter (i.e. representation of the formants) coefficients, a Linear Predictive Coding (LPC) is utilized. The filter coefficients are used to determine the shift in the non-zero imaginary terms of the poles (determined by α) and then converted back to LPC coefficients and together with the residual used to resynthesize the new anonymized speech frame in the time domain. This drafted approach requires neither training nor large amounts of training data. It simply alters the original speech using signal processing techniques to change the voice impressions of the speaker (Sinha and Siegert, 2022).

2.2. Real-time voice changer

A Real-time voice changer is usually a device or software, which can change the impression of a voice by changing the tone or pitch of a voice, adding distortions to the user's voice, or combining the previous methods in various ways. During the current analyses, we relied on Voxal from NHC software offering good usability. Besides the amplification or attenuation of some frequency components, high and low passes are often used as pre-configuration in voice changers. As high-pass filters had a very high fundamental frequency when being applied to female voices, we observed some problems for male voices using the same setup. Consequently, male speakers are very difficult to understand, since important fundamental frequencies are no longer present. The opposite case led to similarly poor results. Thus, to have an anonymization technique that works for several speakers without having to manually tune it individually, a very careful approach in setting the various filter options is necessary.

2.3. TTS-based anonymization

For comparison, we also included a TTS-based anonymization and relied on eSpeak. eSpeak is a compact open-source speech synthesizer. It uses a "formant synthesis" method, allowing many languages to be provided in a small size (Phutak et al., 2019). The speech is clear, and can be used at high speaking rates, but is not as natural or smooth as larger synthesizers which are based on human speech recordings. The advantage of formant synthesis is that it needs less computation and generates reliably intelligent speech output even at very high speeds with small memory footprint for the engine and its voice data and can therefore be easily included in IoT devices without huge performance loss. In the current contribution, the main purpose of the TTS anonymization is to serve as a ground truth to evaluate the anonymization success of the McAdams and real-time voice changer.

3. Experimental Setup

3.1. Dataset

We used spontaneous interactions between humans and a voice-assistant from the "Voice Assistant Conversation Corpus" (VACC), cf. (Siegert, 2020). It consists of

high-quality device-directed and human-directed German spontaneous speech, recorded by 13 male and 14 female speakers of a mean age of 24 ± 3.32 years. The recordings took place in a living room-like surrounding so that the participants could get into a more informal communication atmosphere compared to a laboratory setting. In this analysis, we only used the device-directed speech, which comprise approximately 3,800 utterances with an average length of 2.3s (min: 0.02s and max: 6.09s).

3.2. Anonymization Ability

A pre-trained speaker recognition model (VGGVox) was used to test the identification ability of the anonymized speech samples. It is based on a VGG-M architecture and adapts a deep-CNN architecture, cf. (Nagrani et al., 2017). The model was trained on a large-scale dataset called VoxCeleb1, consisting of over 140,000 utterances by 1,251 celebrities with a wide range of different ages, accents, nationality, etc. Ergo, the model learns speaker-specific cues and prosody mannerisms comprehensively. Furthermore, we use the missrate or false negative rate (FNR), i.e., the number of times the predicted speaker is not the same as the actual speaker. A euclidean distance is used to classify the speaker ID by comparing the speech feature vectors from the test speech, here anonymized speech, with all the enrolled speakers' speech feature vectors. A score of 0 or 1 is assigned according to the speaker ID prediction to refute or confirm the test speaker, respectively.

3.3. ASR Performance

In order to evaluate the degradation in intelligibility caused due to anonymization is measured by Word Error Rate (WER). We used a popular ASR, by Google, that performs well to recognize spontaneous speech, as identified in previous studies (Silber-Varod et al., 2021; Siegert et al., 2020b). The recognized ASR transcription is compared with a reference text to calculate the WER by counting the number of substituted (S), deleted (D) and inserted (I) words against the total number of words (N) in the reference text. In previous experiments, we identified a WER for the original dataset of 0.09 in average (Siegert et al., 2020b).

4. Results

Using device-directed utterances from the 27 speakers of VACC, we generated anonymized speech using: i) McAdams coefficient and ii) eSpeak TTS synthesizer, and iii) Voxal voice changer. In the first method, the McAdams coefficient α is set to 0.8, achieving a good compromise between anonymity and ASR performance, identified in previous experiments (Sinha et al., 2022). Whereas for eSpeak, synthesized speech samples are simply generated by providing the transcription of each speech sample. For Voxal, a specific general configuration consisting of a pitch shifters and amplifiers was utilized. The in such a way altered speech is the anonymized speech, which is then used to further evaluate the success of anonymity and degradation in ASR performance.

Table 1: Anonymization ability and ASR performance of the analyzed techniques.

Anonymization technique	Missrate (in %)	WER
McAdams ($\alpha = 0.8$)	38.72	0.18
eSpeak	83.21	0.68
Voxal	70.11	0.30

We evaluated speaker anonymization and the ASR performance. The overall results regarding ASR intelligibility, measured as WER, and performance in anonymization, measured as missrate, are given in Table 1. Regarding the WER it can be seen that all anonymization techniques result in a worse WER than the original samples. Hereby eSpeak results in a quite high WER and thus a very low ASR intelligibility. The best WER can be observed with the McAdams technique. Regarding male and female speakers, no difference in the WER is observable.

Regarding the missrate, in first sight Voxal and eSpeak result in a better anonymization, but this is due to the much lower ASR intelligibility. When distinguishing male and female speakers it is apparent that for male speakers the missrate is significantly lower than for female speakers, i.e. female speakers achieve a better anonymity, see Fig 2. For the TTS-based system, we are a bit surprised that such a large amount of TTS-voices can fool the automatic speaker verification at all, as the listening impression of the generated voice is quite different from the original sample. Furthermore, it could be assumed that a TTS-voice which is obviously not one of the original speakers will never be identified as one of the known speakers. We assume that this observation is connected to the way the voice samples and the previously trained voiceprints are compared (distance-based similarity measure). The differences in the miss-rate between male and female speakers could be just partly explained by the fact, that we use a male TTS-voice, as also Voxal and the McAdams coefficient show a similar gender bias. It seems as for female speakers a higher anonymization could be achieved. With original, i.e. non-anonymized data, the speaker identification model does not show this behavior. Therefore, we assume that during the anonymization specific female speaker characteristics are changed, while being left unchanged for male speaker. But additional experiments are needed to test thy hypothesis.

5. Conclusion

In this contribution, we discuss several possibilities to allow a fast anonymization without a speaker adaption phase, usable for one-shot speaker anonymization while interacting with public voice assistants. We compared a formant shift algorithm and a real-time voice changer, for comparison we also included a TTS system. To evaluate the success of the anonymization, we measured the

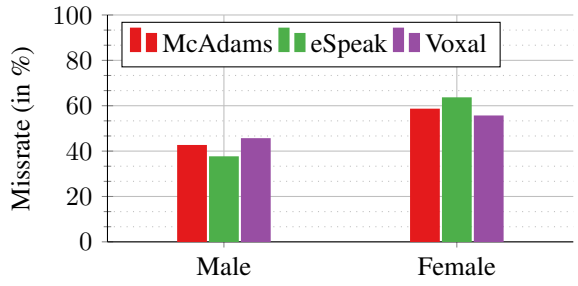

Figure 2: Comparison of miss-rate performance.

anonymization ability using a pre-trained speaker recognition model and the corresponding data. Furthermore, we evaluated the ASR performance using a state-of-the-art cloud based ASR-service. Regarding the anonymization, we could show that for the selected techniques, anonymization ability and ASR performance are connected. With this actual available one-shot anonymization techniques, it is not possible to achieve both a good ASR performance and a good speaker anonymity, so far. This approach maybe suitable for short interactions and many users. Especially, if the speech content itself does not contain personal information, which is often the case for public voice assistant interactions for information purposes, cf. (Kisser and Siegert, 2022).

6. Acknowledgements

This research has been funded by the Federal Ministry of Education and Research of Germany in the project Emonymous (project number S21060A).

7. Bibliographical References

European Court of Justice (Grand Chamber). (2020). Judgment of 16 July 2020, ECLI:EU:C:2020:559. ECLI:EU:C:2020:559.

Kisser, L. and Siegert, I. (2022). Erroneous reactions of voice assistants "in the wild" — first analyses. In *Elektronische Sprachsignalverarbeitung 2022. Tagungsband der 33. Konferenz*, volume 103 of *Studientexte zur Sprachkommunikation*, pages 113–120, Sonderborg, Denmark. TUDpress.

Lopatovska, I. and Oropeza, H. (2018). User interactions with "alexa" in public academic space. *Proceedings of the Association for Information Science and Technology*, (55):309–318.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., and Evans, N. (2019). The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding. In *Proc. Interspeech 2019*, pages 3695–3699.

Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., and Evans, N. (2020). Speaker anonymisation using the mcadams coefficient. *arXiv preprint arXiv:2011.01130*.

Phutak, V., Kamble, R., Gore, S., Alave, M., and Kulkarni, R. (2019). Text to speech conversion using raspberry-pi. *International Journal of Innovative Science and Research Technology*, 4(2).

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice interfaces in everyday life. In *Proc. of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12.

Schönherr, L., Golla, M., Eisenhofer, T., Wiele, J., Kolossa, D., and Holz, T. (2020). Unacceptable, where is my privacy? exploring accidental triggers of smart speakers.

Siegert, I., V.Silber-Varod, Carmi, N., and Kamocki, P. (2020a). Personal data protection and academia: Gdpr issues and multi-modal data-collections "in the wild". *The Online Journal of Applied Knowledge Management: OJAKM*, 8:16 – 31.

Siegert, I., Sinha, Y., Jokisch, O., and Wendemuth, A. (2020b). Recognition performance of selected speech recognition apis – a longitudinal study. In *Speech and Computer*, pages 520–529, Cham. Springer.

Siegert, I., Weißkirchen, N., Krüger, J., Akhtiamov, O., and Wendemuth, A. (2021). Admitting the addressee detection faultiness of voice assistants to improve the activation performance using a continuous learning framework. *Cognitive Systems Research*, 70:65–79.

Siegert, I. (2015). *Emotional and user-specific cues for improved analysis of naturalistic interactions*. Ph.D. thesis, Otto von Guericke University Magdeburg.

Siegert, I. (2020). "Alexa in the wild" – Collecting Unconstrained Conversations with a Modern Voice Assistant in a Public Environment. In *Proc. of the 12th LREC*, pages 608–612, Marseille, France. ELRA.

Silber-Varod, V., Siegert, I., Jokisch, O., Sinha, Y., and Geri, N. (2021). A cross-language study of selected speech recognition systems. *The Online Journal of Applied Knowledge Management: OJAKM*, 9:1 – 15.

Sinha, Y. and Siegert, I. (2022). Performance and quality evaluation of a mcadams speaker anonymization for spontaneous german speech. In *Fortschritte der Akustik - DAGA 2022*, pages 1185–1188, Stuttgart, Germany.

Sinha, Y., Wendemuth, A., and Siegert, I. (2022). Emotion preservation for one-shot speaker anonymization using mcadams. In *Elektronische Sprachsignalverarbeitung 2022*, pages 235–242, Sonderborg, Denmark.

Steven, M., Pan, D., and Engineer, M. (2017). A case study on using voice technology to assist the museum visitor. In *MW17: Museums and the Web 2017*.

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., et al. (2021). The voiceprivacy 2020 challenge: Results and findings. *arXiv preprint arXiv:2109.00648*.

WP29 (Article 29 Data Protection Working Party). (2014). Opinion 05/2014 on anonymisation techniques.