International Speech Communication Association (ISCA)
https://www.isca-speech.org/

ISCA Special Interest Group "*Security and Privacy in Speech Communication*" (SIG-SPSC)
https://www.spsc-sig.org/

RE: EDPB Guidelines 02/2021 on Virtual Voice Assistants
https://edpb.europa.eu/our-work-tools/public-consultations-art-704/2021/guidelines-022021-virtual-voice-assistants_en

Dear Members of the European Data Protection Board,

These are comments to the EDPB Guidelines 02/2021 on Virtual Voice Assistants on behalf of *the International Speech Communication* Association (ISCA) in general and in particular, its special interest group "*Security and Privacy in Speech Communication"* (SIG-SPSC). The association ISCA is the largest international community of researchers in the speech sciences with 2500+ participants in our yearly flagship conference Interspeech. The group SIG-SPSC within ISCA is the largest international interest group devoted to questions of privacy and security in the speech sciences with 120+ members. Your guidelines on virtual voice assistants are thus central to our area of interest and expertise, and we are excited to have the opportunity to share our comments, expertise and opinions with you.

As a general opinion, we are pleased that the EDPB has taken the time to address VVAs, which is an emerging technology where deficiencies in data protection are currently clearly evident and might come to have important and detrimental consequences. It is thus the right time to create VVA guidelines to avoid such consequences. We are also mostly content with the general attitude and extent of the guidelines.

On the top level, our most important comment to the guidelines is related to the way it refers to "unique identifiability" with regard to speech signals. In our view, *we do not have any technology which would provide "unique identifiability" nor do we see any path in the reasonable future which would generate such technology*. The issue is that all speaker identification methods are *probabilistic* in nature, just in the same way as wake-word detection is described as a probabilistic method in the draft guidelines. In fact, all information extracted from speech signals is always probabilistic in nature in the sense that

such information always has a level of inherent uncertainty. This means that the answer from an analysis of a speech signal can be for example of the forms:

- In speaker identification:
    - The most likely identity of the speaker is person A, (and the second most likely identity is B, and the third most...).
    - The likelihood that the speaker is person A is 35.1%, person B is 34.9%, person C is 7.0% and so on.
    - An evidence measure (a measure of evidence for and against) indicates that the speaker is person A (but does not say anything of other persons).
    - It is 10 times more likely that a particular speech signal originates from person A than from another person.
- In emotion detection:
    - The likelihood that the emotion of the speaker is *Happy* is 34 %, *Tired* is 7%, *Angry* is 1%, *Excited* is 70% and so on.
- In gender detection:
    - The likelihood of the speaker's gender is for Male 72%, Female 24%, and Other 4%.
- In speech recognition:
    - The likelihood that the spoken sentence was "How to recognize speech?" is 76% and that it was "How to wreck a nice beach?" is 73% etc.

A consequence of this is that all discussion about information extracted from speech must reflect the level of certainty and uncertainty inherent to that information. Any decisions and inference made based on such information are thus also vulnerable to errors since it is potentially based on incorrect assumptions and premises. On a system level, VVAs must therefore be designed to be robust to errors in interpretation and risk-averse in their actions especially with respect to private information and real-life consequences.

Importantly, the certainty of information depends very much on the context. For example, if we are to identify a person from a small group of people, like those employed at a particular small company, then the task is easy and the result will have high certainty. Should we on the other hand want to uniquely identify or verify the identity of a person among all people of the world, then the result will obviously have very high uncertainty.

Further consequences include that when the speaker identity always involves uncertainty, it becomes difficult to determine whether the current speaker has already consented to the processing of private information. In fact, speaker recognition itself involves analysis of private biometric information such that determining whether a user has a contract with the service provider already potentially reveals biometric information to the service provider. A VVA must therefore be risk-averse in its voice biometrics, such that voice biometrics is invoked only when it is explicitly required.

The above-mentioned ambiguity in terminology with respect to "identifiability" is especially evident when the guidelines describe the technology of speaker recognition as being based on a "speaker template". We are not aware of any state-of-the-art speaker recognition algorithm based on a concept of such a "template". Instead, speaker recognition methods are based on probabilistic models which describe the range, variance and combinations of

voice features that a particular speaker typically exhibits. The difference is not only semantic but reflects an important difference in the way of thinking about voice biometrics from a probabilistic perspective as described above. While computers need to store fixed-length data to represent an identity that is to be associated in voice biometrics with a new voice sample, this fixed-length serialisation is not necessarily a template. Throughout a speech algorithm, it is used inherently to parameterise a probabilistic interpretation.

A consequence of probabilistic speech representation is also the inherent need to address uncertainty in decision outcomes and to limit the impact of uncertainty in decision making. Thus, precision (used here as the mathematical opposite of uncertainty) cannot all and always result in some divine uniqueness. Moreover, any notion of "beyond reasonable doubt" is pointless in speech technology - on the contrary: given a specific level of reasonable doubt, especially in settings where decisions must be made to ensure a positive user experience, a decision will be made that reflects a low-risk impact in the light of cost assumptions arising from the impact of all possible decision outcomes (that are reasonable to pursue within a frame of decision logic).

Our second main comment is that among the stakeholders, it is not only the service providers who can have access to information but also other users. As a practical example;

> A current VVA service provides a voice command history, where the user can listen to past commands to the VVA and that command history can also be accessed through any of the devices connected to the same account and service. As a consequence, a user can access the command history of his home-device even when at the office or travelling, *even when the commands have been made by other users.* In other words, users have by default access to each others' voice command history. For example, partners and room-mates can thus covertly monitor each others' voice commands. This does not adhere to users' typical expectations of privacy.

The guidelines should therefore contain a section that gives recommendations on how to handle privacy when there are multiple users of a single device. For example, the recommendation could be that user A would have access only to information related to interactions where the same particular user A has been identified. In particular, user A would by default *not* have access to information from interactions of a user B where A was not present, nor to interactions where the speaker has not been identified. Voice command history would also be accessible only for users who have been positively identified with sufficient certainty. Access to other users' information should be given only through explicit consent (opt-in) of those other users. Conversely, storage of voice command history might be allowable only when all users of an interaction have been positively identified with sufficient certainty.

A third point is that the guidelines inaccurately states frequently that VVAs are *all* based on cloud services to some extent. There are, for example, open-source VVAs available that can be implemented on a local device without Internet access, such as a Raspberry PI. Clearly, there are benefits for the privacy of the user from such devices even if the functionality and performance are necessarily constrained. The guidelines should better take into account that privacy can be preserved much better with such devices and that cloud services are not mandatory for all functionalities.

Finally, we have made a collection of more detailed comments from our community to the document listed below. Though we have tried to be brief, the list of comments is long and we are happy to provide more details and commentary on request.

on behalf of

the Special Interest Group "Security and Privacy in Speech Communication" (ISCA SIG-SPSC) and the International Speech Communication Association (ISCA)

Tom Bäckström, chair of ISCA SIG-SPSC
associate professor at Aalto University, Finland

Andreas Nautsch, secretary of ISCA SIG-SPSC
senior research fellow at EURECOM, France

Specific comments

- Page 9; Sections 2.3 and 2.4; We believe that the wake-up word procedure which is nicely described in Sections 2.3/2.4 lacks privacy/security mechanisms in case the audio is passed to the cloud-based server for a second check on the wake-up word. In this case, the recommendations should include the use of privacy-enhancing technologies (homomorphic encryption, data minimization and the like) to make sure that collected audio is used for the sole purpose of wake-up word detection.
  Another option would be that the local device gives audio feedback without communication with the cloud when it has detected a wake word with low confidence. That is, if the device identifies something which is similar to a wake word, it could play a "question mark" sound, "Huh?", such that the user would notice that the device is confused. If the user indeed had said a wake word, then the user could say it again, otherwise, the user would just be warned that the wake word was almost triggered.

- Page 9; Sections 2.3 and 2.4; This report relies on the concept of wake words. Some recent industry news suggests that certain VSS may relax the wake-word concept and move to other means for invoking the VSS
  https://www.theverge.com/2020/10/22/21528133/google-nest-hub-smart-display-proximity-wake-word-ultrasound
  https://www.amazon.science/blog/alexa-do-i-need-to-use-your-wake-word-how-about-now
  For the development of these guidelines, It would be mindful not to depend exclusively on the existence of a wake word.

- Page 10 point 18; It is a good observation that wake-word detection is "probabilistic". In fact, all functionalities of VVAs are probabilistic. It is important especially in speaker identification because a VVA can never find the "unique identity" of a speaker, but only a probability of who the speaker could be. The probabilistic nature of speaker identification should be more pronounced throughout the guidelines. A discussion on terminology, in particular on the different meanings of "identification" and "unique identification" can also be found in

  Jasserand: "Legal Nature of Biometric Data: From 'Generic' Personal Data to Sensitive Data", European Data Protection Law Review, 2(3), 2018, pp. 297-311
  https://edpl.lexxion.eu/article/EDPL/2016/3/6

- There seem to be some ambiguities or even contradictions in the document. In paragraph 117 it is stated that "If working as currently designed, VVAs do not send any information to the speech recognition cloud service until the wake-up expression is detected." which is vague or even contradicts earlier observations in paragraph 19 that data is sent to a server for a second round of the wake-up word detection mechanism. This should be clarified in the sense that these devices do not send any information until it is likely that the wake-up word could have been spoken (leaving room for interpretation).

- p.12 point 29: "Conversely, such consent as required by Article 5(3) e-Privacy Directive would be necessary for the storing or gaining of access to information for any purpose other than executing users' request (e.g. user profiling)."
  -> Clarify that, besides user profiling, this also includes annotating and using the data to train an ASR/NLU system. This is clarified later on p.22 but mentioning only user profiling here could be misleading.

- Page 12 point 29; "Data controllers would need to attribute consent to specific users. Consequently, data controllers should only process non-registered users data to execute their requests." -> This is unclear; it would be helpful to further explain this.

- Page 12 point 31; Highlight also other features than only sex or age, such as "mood", "attitude". The spoken content also carries additional metadata - education, regional dialect, etc.

- Page 13 section 3.2.1; Should add a subsection with a discussion about the probabilistic nature of information; Specifically, all information extracted from speech signals is always probabilistic in the sense that we never have full certainty of the content. Instead, for example, 1) we can say that "the most likely" content is X or 2) we can say that the content is X with a probability of Y percent, or 3) with an evidence measure indicating that the content is X.
  Consequently, all processing must take into account that the assumptions and premises of any decisions and inference can always be incorrect, such that also the made decisions and inference are incorrect.

- Page 13 section 3.2.1; Also other users can sometimes access voice history, so they can "process" information. Access management between users is needed.

- Page 16 point 49; There's a looming issue with consumer protection ahead: multi-user, data ecosystem, interface specificities - if a household has multiple VVAs, information overload is simply unavoidable. Moreover, if changes are occurring in any settings of a VVA regarding multi-user, data ecosystem, interface specificities - providing information to data subjects is one thing; these being actually informative is another (are end-users truly empowered?). There's a strong need for pictograms, icons, and symbols. Compare with ISO/IEC 24779-1:2016 for use with biometric systems (this one is written for border control mainly; it is provided here purely informative to the EDPB to indicate the necessary work ahead). Transparency is in high demand to also increase consumer protection.

- Page 17 point 54; All VVAs do not require registering a user.

- Page 8 Section 2.2; It's useful to have a list and definition of roles in the VVA ecosystem. Should accidental users be mentioned separately?

- Page 18 point 55: It seems likely that voice control can currently be used to control devices and services without violating the transparency principle of GDPR. The implications seem far-reaching: Is it then possible to control third-party services and products with voice assistants without violating the GDPR? Does the GDPR

constrain or even block the use of IoT?
At the same time, the guidelines also point out that if it is necessary for the performance of a request (e.g. command Turn on the light), users do not have to be asked for their explicit consent. We would be interested in whether this can be used to solve the transparency problem, i.e. only exchange data between devices when explicitly requested and then delete data after processing.
We hope that you can solve the transparency condition in such a way that the usability aspect of the VVAs is not impaired in the long term. If explicit consent is required for every request, it will surely be a show stopper for many users and applications.
Could the problem be solved in such a way that explicit consent must be given when a certain VVA service is called up for the first time, and this consent could then be dispensed with for further use? If a service is not used for a longer period of time, the associated transparency status is then reset, so that when it is used again, explicit consent is required again.
This suggestion is inspired by the current Google Android OS: Here the permissions given by the user to an app (e.g. activation of the microphone) are withdrawn after a long period of non-use.

- Page 18 point 56: With regard to the inclusiveness of VVAs, it is important that all information about the user privacy rights are available in the same modality as it is accessed. For example, a blind person cannot read a privacy statement, but he/she can use a VVA. Since anyone using a VVA necessarily can speak and hear, privacy rights must therefore be available at least in spoken form.
  It is further helpful if the same information is available also in a textual digital form since that enables easier searching and cross-referencing the document than an audio format.

- Page 20; Need to add commentary about transparency with respect to access for other users. What information of user A can other users B,C, and D access?

- Page 21; Public VVAs (such as an information desk) do not need registration; Shouldn't there be an information sign on such public assistants, similar to the video surveillance perhaps? Visual information is however problematic from an inclusiveness point of view because blind people can use the VVA but not notice the information sign.

- Page 21; A contractual question; Can a primary user (owner) of a VVA device accept a contract that covers other users of the VVA? Is that necessarily a liability for the primary user, such that by accepting the contract the owner is likely to cause legal problems for him/herself? Is it plausible or practical that an owner of a VVA at his home/office will explain the privacy-level to all guests arriving? Is it then ethical, and does it follow the spirit of the GDPR to even give that as an option to the owner?

- Page 21; Any contract and consent to be valid, the VVA must have speaker verification (biometric information or otherwise) with high confidence.

- Page 22 footnote 30; the guideline discusses here biometric terminology without referencing the harmonised biometric vocabulary standard (ISO/IEC 2382-37:2017) while saying almost exactly the same. Furthermore, terminology clash is ahead when terminology is not qualified adequately: instead of "verification", one could refer to "biometric verification". To distinguish between/trying to map vocabulary among different involved communities is good anticipation (e.g., "verification" from biometrics and "authentication" from IT-security), yet, a glossary might be a more appropriate format for doing so (this would also be a fantastic place to disambiguate between contextual nuances of each community when using their terminology in their domain - just think about "discrimination power" which is a good term in machine learning but devastating in any social study field). The EDPB should maybe not refer to authentication at all.

- Page 26 point 102; "acceptable threshold" is vague. At least some criteria for the threshold would be important.

- Page 26, point 105; The co-organiser team of the VoicePrivacy initiative themselves struggle with how to use terminology coherently as regards what "anonymisation" and what "pseudonymisation" mean (there are different viewpoints, but when no consensus can be found, one needs to pragmatically move on). The academic aspiration naturally is to do as good as possible (anonymisation), yet, data resulting from human interaction is not just some sterile computational alphabet that is susceptible to subsequent theoretic proofs (like zero-knowledge proofs in cryptography). On empirical data, trying to remove sensitive factors to achieve data privacy can only result in levels of "pseudonymisation". If some _perfect result_ is demonstrated for a particular dataset, it might be for that particular dataset only (empirical proofs are no bulletproof guarantee like a theoretic proof; yet, theoretic proofs are unavailable with speech data). Are the terms ``anonymisation" and "pseudonymisation" in this EDPB guideline used as they do understand their usage in the GDPR? There is serious doubt, whether or not voices can be anonymised - would the end-result still be a voice? How is a "voice" or "speech" defined in these guidelines? Is the use of these words consistent?

- p.26 point 105: The two articles cited in the footnote are irrelevant. The paper by Cohen-Hadria et al. does not "remove situational information like background noises". On the contrary, it aims to preserve background noise and destroy any overlapping speech. The paper by Qian et al. provides almost no protection as shown in https://hal.inria.fr/hal-02355115/file/ppvc_final.pdf. The VoicePrivacy Challenge baseline provides much better protection.

- Page 26 point 107; "render voice unidentifiable", "anonymization" does not reflect technological capabilities. Would benefit from using the same vocabulary and standards as the biometric and forensic communities. Particularly, the word "unidentifiable" raises concerns. The border control biometrics community (e.g., European Association for Biometrics) and the voice biometrics community (see discussions after Els Kindt's keynote at the 2018 Speaker Odyssey workshop) criticise the GDPR wording "unique identification" which provides no statistical control whatsoever; there are even different legal interpretations on its use (100%

match without errors - which is impossible; shortlisting in a set of potential identities - even if the true speaker identity is not enlisted before). Similarly, what does "unidentifiable" mean? While industrial systems estimate, whether or not some specific confidence level is met, voice biometrics treats so-called "likelihood ratios" which put two competing propositions in a tug-of-war situation (one might compare it to prosecutor and defendant are presenting evidence in front of a judge/jury; whoever is stronger, the decision will be made for). The European Network of Forensic Science Institutes (ENFSI) provides categorical tags for such likelihood ratios when forensic practitioners are reporting evidence in court: Willis et al. "ENFSI guideline for evaluative reporting in forensic science", ENFSI, 2016. Motivated by the ENFSI guideline, a metric for pseudonymised human data has been proposed in

Nautsch et al.: "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment", Interspeech 2020, p. 1698-1702
http://dx.doi.org/10.21437/Interspeech.2020-1815

N.B.: the forensic community is undergoing a paradigm shift since decades regarding the strength of evidence in their lab method validation. Without a definition of "voice", "voice data", "voice representations" in the first place, no statistical control can be provided, and by consequence, the guideline wording will become meaningless.

- Page 27 point 109; The GDPR is general and while it is understood that it is a reference to this guideline, it could help to indicate here which part of the GDPR is aimed at in which way (e.g., examples with interaction in communication between company & DPOs if that is meant here). Is this point referring to company-internal reviews or to communication to DPOs using DPIAs? Instructions need to be clear; there is a lack of precision here. Data controllers have the obligation to periodically review their data processing operations; how does this depend on internal procedures - and internal to whom?

- Page 27 point 110; Speech is important for human development, whereas the guideline presents it as "creates a new set of security risks" (speech/other means of communication are essential to newborns; see language deprivation experiments, such as Frederick's experiment; speech is vital, fatal to newborns if abstinent, not some security risk). While the context of the guidelines is clear, it really surfaces here that "voice" needs to be defined at the beginning of the document, which it was not. As the example might illustrate: there are different perceptions of "communication means" as well. Please define your perspectives, and explain their meanings.

- Page 27, points 112, 114, 127, 129, 134; the comparison between "fingerprint" or "voiceprint" is tempting but inherently misleading. A particular fingerprint is something fixed that allows us to observe the same pattern over and over again. Voice/speech/communication is always in flux and nothing is static.
  The term ‘voiceprint’ can be found in several foreign laws as biometric identifiers (e.g. in the Illinois Biometric Information Act and in the Texas Business and Code of Commerce). However, several researchers have challenged the use of the term as it

gives the impression that a voice can be ‘graphically’ represented. See for instance Boë who writes that voiceprint is ‘an erroneous metaphoric terminology [that] leads many people (not only the general public) to believe that a graphical representation of the voice (sonogram, as it happens) is just as reliable as the structure of the papillary ridges of the fingerprints, or genetic fingerprints, that it allows reliable identification of the original speaker’ (Boë, Forensic voice identification in France, Speech Communication, 2000, 31, 205). See this analysis made in Nautsch et al.: "Preserving privacy in speaker and speech characterisation", Computer Speech & Language, vol. 58, 2019, pp. 441-480 - page 448. https://doi.org/10.1016/j.csl.2019.06.001

On the use of technical terminology in general, we would recommend the EDPB to refer to the ISO/IEC Standard on Harmonized Biometric Vocabulary (ISO/IEC 2832-37, (“Information Technology – Vocabulary – Part 37: Biometrics”). Pre-GDPR, the Italian Data Protection Authority (Il Garante) acknowledged the authority of the standard for the technical definitions in its ‘Guidelines on Biometric Recognition and Graphometric Signature, Annex A to the Garante’s Order of 12 November 2014.’
https://www.garanteprivacy.it/documents/10160/0/GUIDELINES+ON+BIOMETRIC+RECOGNITION.pdf/3ac0d4ff-7575-4f5e-a3fa-b894ab7cf517?version=1.1

While the GDPR has introduced a legal definition of biometric data, this does not replace the existing technical definitions.
On terminology, see the collaboration between the technical and legal communities in
Nautsch et al.: "The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps towards a Common Understanding", Interspeech 2019, pp. 3695-3699 http://dx.doi.org/10.21437/Interspeech.2019-2647

- Page 30, points 127, 128, 129, 131; the term "template" should be abandoned, when talking about speech; it does not reflect the nature of any speech technology. The standard for a biometric voice data interchange format (ISO/IEC 19794-13:2018) refers to the more general term "voice representation"; this simply leaves any co-annotation open (although it concerns rather raw audio data). In specific text-dependent settings, where end-users repeat the same passphrase over and over again, the thought construct of a "template" may suffice, such as some sorts of personalised wakeup word detection. For text-independent technology, this is not the case at all. Compare it to posting the same tweet all over again to the capacities demonstrated by natural language processing in the Cambridge Analytica scandal.

- Page 30 point 128. “In the example above, the recognition of a the user’s voice” -> “In the example above, the recognition of a user’s voice” (delete the)

- Page 30 point 132; Perhaps it is better to simply say that the activation should be under the control of the users themselves. Example 12 makes a mix-up giving the impression that authentication is different from verification. Moreover, elaborating on the suggestion that multiple VVAs in one household repeatedly ask "do you wish to be identified" - it could be a bit too much (long-term, also).

- Page 31, point 134; Why not simply referencing the harmonised biometric vocabulary standard? The ISO/IEC 24745 standard is on "Biometric information protection" - not on template protection as written in the guideline (this is wrong). There is more data across the stages of the biometric processing pipeline which contain biometric information than on the template level alone. Also, the ISO/IEC 24745 standard defines what measures for unlinkability, renewability, and irreversibility should look like in principle, but no definition of such metrics is provided (without statistical control, 24745 facilitates well-feeling comfort only). Moreover, during ISO/IEC JTC1/SC37 meetings and expert discussions therein, it became clear that the disambiguation of when one concerns "biometric features" and when one concerns "biometric templates" depends on system integrations (not on the technology that was to be integrated). Although two integrations might use the same technology, their particular setting can lead to different considerations when the same numerical data is referred to as "feature" or "template", or something else entirely such as a "(probabilistic) model". Drawing concise boundaries will not be the purpose of this guideline; please use & define more general terms such as "voice representation", "voice data" or "voice information". Is there a difference between "voice" and "speech" and "verbal human communication" or are all these interchangeably alike?

- Pages 32-33, points 147 and 150; Unfortunately, one ongoing development in the IoT area is that the large IT companies are building their own isolated IoT ecosystems and are not really interested in universal networking. With regard to GDPR, this could even have advantages, since the legal responsibility for the provider and user is less complex. However, it is also repeatedly criticized that the verticalization of IT infrastructures leads to a strong distortion of competition and that users cannot easily switch to another service (together with their profile data). It should therefore be achieved that the implementation of the GDPR requirements for VVA does not contribute to the fragmentation of IoT services, but counteract it (somehow supported by point 171 and 172, p. 37f).
  Then the question arises of how data portability and accountability can be reconciled. Imagine you control devices via Alexa. Then you decide to switch to Google Assistant. However, some of your devices are not compatible with Google Assistant because the device developers have contracts with Amazon. You can not port your data because of the business decisions of Amazon and the device developers. However, for users, it may appear that the problem lies with Google because it does not support the switch. Should Amazon or the device developers be obliged, if they restrict data portability, instead of data to prepare messages for the users which make it transparent that they do not support Google? For example, a text file that the Google Assistant can read like "Sorry, the data from device X can only be read by Amazon Alexa devices ..."

- Page 33 point 152; “VVA designers, as well as app developers in case they are part of the solution, should at the end of the exercise process inform the user that his/her rights have been duly factored, by voice or by providing a writing notification to the user’s mobile, account or any other mean chosen by the user.” -> Meaning is unclear. It would be useful to get a further explanation here.

- Page 34 point 156; the assumption shines through that all VVAs must store speech data and its derivatives (e.g., transcriptions) centralised (which is not accurate). Is it DPIA depending on whether or not personal data (besides audio/voice data) would be centralised, whereas audio/visual voice data could be on a device? What is the intention here?

Other comments not related to specific sections of the guidelines:

- Speech/voice might be beyond acoustic data only, such as audiovisual data or text-form (e.g., through automatic transcription). Would the Microsoft Kinect for Xbox One qualify as a VVA in non-/gaming applications? It captures video & audio, and offers speech recognition modules for several languages, and also skeletal tracking. Audiovisual word recognition has been successfully demonstrated in the EU project 706668 "TalkingHeads", a "Marie Curie" individual fellowship in the domain of Audio-Visual speech recognition.

- Speech technology might have trouble distinguishing between twins. In forensic settings, when (telephone) bandwidth is much more limited than for VVAs, this is referred to as the "brother effect" (they just sound so alike over the phone). We like to point you to the conclusion of the following article

  H.J. Künzel: "Automatic Speaker Recognition of Identical Twins", International Journal of Speech Language and the Law, 17(2), 2010, pp. 251-277 https://dx.doi.org/10.1558/ijsll.v17i2.251

  "Speaker identification of identical twins will never become a standard task in the forensic environment, the most trivial reason being the low incidence of one in 250 births in Western Europe (Dudenhausen 2003: 301). However, the extreme form of genetic similarity that also extends to the structures used for speech production may help develop a benchmark for the performance of speaker recognition methods, especially automatic SPID systems. A system that identifies an identical twin without falsely accepting the other twin is probably fit for use in the forensic environment."

- User should be aware that privacy is not alike to end-user license agreements (EULAs) - also Common Creatives licenses are not privacy consent - it is not only to inform users about that data is processed but also for the use case and reasons of why which processing. To this extent, the user should have full control/choice. This virtually corresponds to that companies must adapt VVAs for specific international legislation, e.g., in some states biometric processing is prohibited. Thus, products and services must already be designed to a particular level of granularity for companies operating at a global scale. Empowering VVA users for their choices means putting consumer protection compassionately at the same level as data protection. A guideline point could make clear that deliberately giving the choice of activated modules throughout product design and underlying processes must be with the user, provided as quality-of-service by the company (quality as means of

data protection that empowers VVA users and by consequence, consumer protection).

- It may be useful to explicitly name “conversational privacy” as the technical term for (1) informing users about privacy and (2) giving users control over their data in a conversational form. Possible references for this include

  H. Harkous, K. Fawaz, K. G. Shin, K. Aberer, Pribots: Conversational privacy with chatbots, in: Twelfth Symposium on Usable Privacy and Security (SOUPS 2016), USENIX Association, Denver, CO, 2016.
  https://www.usenix.org/conference/soups2016/workshop-program/wfpn/presentation/harkous.

  S. Sannon, B. Stoll, D. DiFranzo, M. F. Jung, N. N. Bazarova, “I just shared your responses”: extending communication privacy management theory to interactions with conversational agents, Proc. ACM Hum.-Comput. Interact. 4 (2020).
  https://doi.org/10.1145/3375188. doi:10.1145/3375188.

  B. Brüggemeier, Conversational Privacy – Communicating Privacy and Security in Conversational User Interfaces (2020).
  https://mediaserver.eurecom.fr/permalink/v125f76418ddcf52uewo/iframe/

  Giving references would enable designers to look up existing information on how to communicate privacy information in a conversational form.

- We find it alarming that no references are provided to any of the speech community's journals or conferences. It reveals that the guidelines do not reflect an understanding of the underlying technology. Some examples of publications:

  - https://www.isca-speech.org/archive/
  - ISCA's main event "Interspeech" features specific tracks to security & privacy, among others, from the SIG-SPSC organised tracks:
    - https://www.isca-speech.org/archive/Interspeech_2019/#Privacy%20in%20Speech%20and%20Audio%20Interfaces
    - https://www.isca-speech.org/archive/Interspeech_2020/#Privacy%20and%20Security%20in%20Speech%20Communication
  - https://www.journals.elsevier.com/speech-communication/
  - https://signalprocessingsociety.org/publications-resources/ieeeacm-transactions-audio-speech-and-language-processing
  - https://www.journals.elsevier.com/computer-speech-and-language/
  - https://asmp-eurasipjournals.springeropen.com/