INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Sound Privacy: A Conversational Speech Corpus for Quantifying the Experience of Privacy

Pablo Pérez Zarazaga[1], *Sneha Das*[1], *Tom Bäckström*[1], *V. V. Vidyadhara Raju*[2], *Anil Kumar Vuppala*[2]

[1]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[2]International Institute of Information Technology (IIIT), Hyderabad, India

(pablo.perezzarazaga, sneha.das, tom.backstrom)@aalto.fi,
vishnu.raju@research.iiit.ac.in, anil.vuppala@iiit.ac.in

Abstract

With the growing popularity of social networks, cloud services and online applications, people are becoming concerned about the way companies store their data and the ways in which the data can be applied. Privacy with devices and services operated by the voice are of particular interest. To enable studies in privacy, this paper presents a database which quantifies the experience of privacy users have in spoken communication. We focus on the effect of the acoustic environment on that perception of privacy. Speech signals are recorded in scenarios simulating real-life situations, where the acoustic environment has an effect on the experience of privacy. The acoustic data is complemented with measures of the speakers' experience of privacy, recorded using a questionnaire. The presented corpus enables studies in how acoustic environments affect peoples' experience of privacy, which in turn, can be used to develop speech operated applications which are respectful of their right to privacy.
Index Terms: Experience of privacy, speech interfaces, speech corpus, acoustic environment, right to privacy

1. Introduction

Since the release of the general data protection regulations (GDPR) in the EU [1], user-privacy is in vogue and the awareness surrounding personal data and its distribution has increased [2]. People have started being conscious about their privacy in online services, as well as the ways their data is handled by digital devices. However, few consider the level of privacy in spoken interactions with digital devices, even though speech is their main mode of communication [3].

Privacy related to speech signals has been considered in studies of room acoustics, where it is possible to limit propagation of the voice by changing the acoustic properties of the room [4–7]. This is especially useful in the design of open plan offices, since people demonstrate a higher productivity when they cannot understand background speech in contiguous spaces [6]. Measures, like the articulation index (AI) and the speech transmission index (STI), have been used to measure the level of privacy of open office environments [4]. In contrast, public venues like theaters and auditoriums require efficient propagation of the voice to reach the audience [5].

According to our personal experience, people intuitively modify the way they talk in a conversation, depending on 1. the content, 2. the level of trust they have in their conversation partners and, 3. the environment, which may condition the amount of information shared. Speech is so strongly rooted in people's habits that they automatically adjust to the environment's privacy requirements. For example, when our conversation might be overheard, see Fig. 1, we intuitively lower our voice level. If this perception of privacy could be measured by our devices, they could also adapt the shared information to the level of trust between devices and the communication environment.

Figure 1: *Malicious person eavesdropping on a conversation.*

Previous attempts to quantify privacy require measurement of the acoustic properties of the room in controlled conditions and using specialized equipment [4]. In contrast, our long-term objective is to develop methods for average and low-cost devices in real-life scenarios. Our aim is furthermore to quantify peoples' perception of real-life environments, where we do not yet know which properties of those environments have an influence on their experience of privacy. To ensure that our recordings reflect such real-life scenarios and the subjects' perception correspond to real-life experiences, we have chosen to record speech in real-world environments using average personal devices. While this choice obviously reduces signal quality and introduces noise in the acoustic environment which we cannot control, it is the most realistic scenario we could construct.

This paper presents a corpus which quantifies how people perceive their privacy in different scenarios and evaluates the acoustic features that define this perception. The recording sessions and the acoustic environments are described in Section 2. The recordings were complemented with an onsite questionnaire which is discussed in Section 3. Additionally, we used an online questionnaire to validate the corpus in terms of the perception of privacy as presented in Section 4. Finally, the structure of the corpus is described in Section 5.

2. Recordings

Our objective is to collect samples of speech which display different levels of privacy. Prior works have shown that the level of privacy can be measured with several different parameters depending on the purpose of the room [4]. Such measurements focus on the speech intelligibility with the objective of 1. improving the productivity of workers, which depends on the background speech that they can hear, or 2. making sure that a theatre audience can understand what the actors say in every seat.

http://dx.doi.org/10.21437/Interspeech.2019-1172

Table 1: *Questions asked to evaluate the perception of privacy.*

Q1	How likely are you to share a secret, in normal voice, in this acoustic environment?
Q2	If you share a secret in this acoustic environment, with what loudness will you be comfortable to share it?
Q3	In this acoustic environment, how likely is it for an eavesdropper to hear your normal voice, in your opinion?

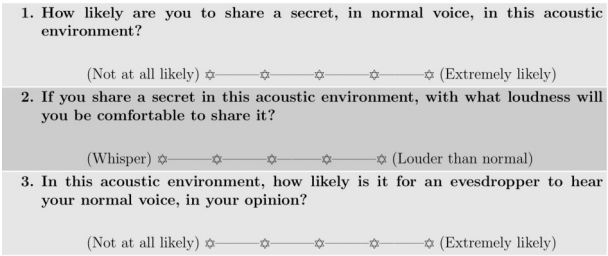

Figure 2: *Screenshot of the questionnaire.*

In contrast, here we focus on the effect that the environment has on the *experience and perception of privacy* of the speaker.

Specifically, our focus is to evaluate how people experience the level of privacy present in different environments depending on the acoustic properties. On one hand, privacy becomes an issue when speaking at public places, which most typically occurs in conversations among people who know each other. On the other hand, we expect that the content of the conversation or dialogue would have a large effect on the impression of privacy – topics related to intimate secrets are more sensitive. To extract the most useful information, we therefore chose to 1. record scripted nonsense dialogues, 2. in different acoustic environments and 3. ask the subjects to evaluate the level of privacy that they perceived in each scenario.

For the best match with real-life scenarios, we recorded speech using common mobile devices, such as One Plus 5T, Motorola Moto E6 and Moto G4 Plus. While the audio quality that these devices provide is lower than that of high-fidelity measurement equipment, they do reflect typical real-life scenarios. Similarly, to ensure that our recordings reflect realistic conditions, we chose to record in real-life environments, instead of simulated ones. Recordings were performed in two geographical locations, at Aalto University in Espoo, Finland and at the International Institute of Information Technology, in Hyderabad, India.

The different scenarios represent typical environments where a conversation would be held, each of them with different acoustic properties and types of noise. The first two scenarios take place in an office with, respectively, the door closed and open, slightly modifying the properties of the environment but maintaining a dry sound due to a short reverberation time. The third scenario is a crowded cafeteria with a strong level of babble and cutlery noise from the people around the conversation. The fourth scenario takes place in a big hall with a long reverberation time. Finally, several conversations were recorded outdoors. The recording devices were placed on a table between the speakers in all indoors scenarios and respectively, outdoors the device was held in hand between both speakers.

The environments in Finland were a $3 \times 4\,\mathrm{m}$ office with two walls made of glass and two made of brick, a cafeteria of $20 \times 15\,\mathrm{m}$ and a big hall of a triangular shape of $50 \times 30\,\mathrm{m}$. The outdoor recordings were collected in a park and next to a road with cars passing by.

In India, the office room was of size $9 \times 5\,\mathrm{m}$, two of its walls are made of glass and the other two are brick walls, the cafeteria was of size $18 \times 15\,\mathrm{m}$, and the big hall has a size of $30 \times 20\,\mathrm{m}$. The outdoor recordings were collected outside the main entrance of the institute, next to a road with heavy traffic.

In each environment, subjects simulated a normal conversation. To increase the fluency of the dialogue and to avoid an impact of the conversation context, a scripted dialogue was generated using random sentences from the Harvard dataset [8]. Subjects were instructed to speak at a normal level that would take place in the corresponding environment.

Most mobile phones sold today include automatic noise reduction algorithms that are applied when any sound is recorded. It was observed that using a recording app [9], noise reduction was only applied in stereo settings and after 45 s of recording. Therefore, the recording settings were set to mono sound and the amount of sentences of every dialogue was chosen so that the length of the whole recording was between 45 s and 1 min.

In total, the voice of 25 speakers distributed in pairs was collected in Finland and 60 speakers took part in the recordings in India. Long speech samples would be difficult to use in listening test scenarios, while shorter samples might not contain enough information about the acoustic environment. The length of 15 s was therefore chosen as a compromise between applicability in listening tests and amount of information for analysis. The recorded files were thus cut to contain blocks of 4 sentences with an approximate duration of 15 s. The resulting audio segments are classified and stored in a public speech corpus.

We designed two questionnaires to evaluate how the subjects perceive the privacy of every scenario. One of the questionnaires takes place during the recording session where every speaker quantifies different features related to the privacy of the environment. To disassociate the perception of the acoustic environment from the visual feedback of the room, another questionnaire is presented as an online form where participants grade similar features only based on the recorded audio.

3. Onsite questionnaire and responses

To quantify the effect of the acoustic space on the experience of privacy of an individual, the participants were prompted to answer a questionnaire immediately after completion of the recording in each acoustic scenario. The objective of the questionnaire is to annotate the audio-recordings with the participants' experience of privacy for the different acoustic scenarios.

3.1. Questionnaire structure

Due to the subjectivity of the perception of privacy, we expected a level of difficulty in answering a direct question on how people felt about the privacy of their surroundings. Thus, to aid the process, the questionnaire comprised of 3 questions, shown in Table 1, designed to be semantically complementary. The analyses of the responses in the latter part of this section will validate if the three questions provide a measure of the same quantity: experience of privacy.

Question 1 (Q1), which we also refer to as the primary question, provides a direct measure of privacy of the acoustic space. In contrast, questions 2 (Q2) and 3 (Q3) indirectly infer the measure of privacy. We expect that the responses to Q3 will be inversely proportional to the feeling of privacy and the responses to Q2 are directly proportional to the measure of privacy.

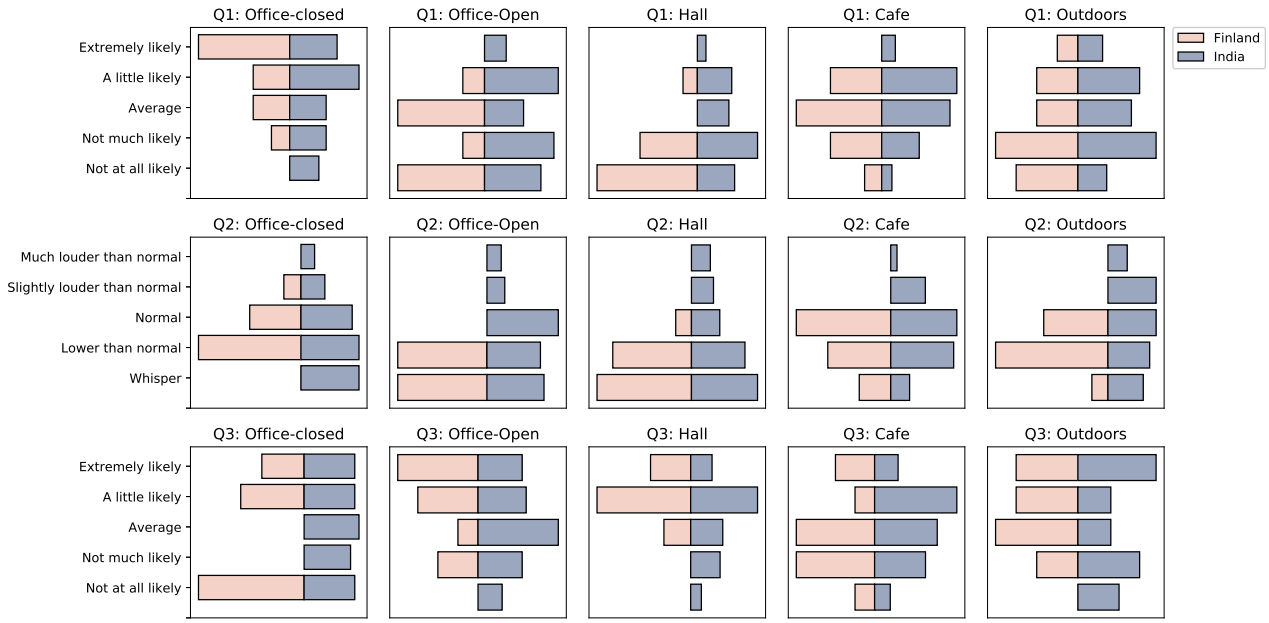

Figure 3: *Results from the onsite-questionnaire conducted in Finland and India to questions Q1, Q2 and Q3.*

The responses are graded on a Likert scale [10] and we employed the commonly used five-level scale for each question. The options for Q1 and Q3 were as follows: 1. *Not at all likely*, 2. *Not much likely*, 3. *Average*, 4. *A little likely*, 5. *Extremely*. The options for Q2 were: 1. *Whisper*, 2. *Lower than normal*, 3. *Normal*, 4. *Slightly louder than Normal*, 5. *Much louder than normal*. A screen-shot of the questionnaire is shown in Fig. 2.

3.2. Hypotheses and responses

The recording scenarios were designed under the hypothesis that 1. the ratio between stationary and non-stationary background noises, and 2. the reverberation of the acoustic environment are the two main criteria in shaping the experience of privacy. In this section we seek to find a preliminary validation of this hypothesis.

Among the five scenarios in Finland, hall, cafeteria and outdoors have mainly non-stationary background noise. The cafeteria-scenario comprises of the most non-stationary background, followed by the hall, which is a mix of non-stationary noise and reverberation. The characteristics of outdoors-scenario in Finland vary depending on the traffic. In contrast, the hall and the outdoors in India are the most non-stationary environments. We expect that the presence of non-stationary noise increases the feeling of privacy. However, this perception could depend on the kind of non-stationary noise. If the non-stationarity arises from clearly audible voices from external speakers, the acoustic space could become less private. In this paper we have not explicitly explored different types of non-stationary environments and it remains a topic for future work.

A highly reverberant and noiseless scenario gives a feeling of being easily overheard, and thus feels less private. While both office scenarios in Finland and India contain no background noise, they vary in terms of their reverberation. We expect that participants would find the office with closed door more private and any difference in the perceptions is due to reverberation being different between scenarios. In Finland, the hall is highly reverberant with some background noise. This is similar to the cafeteria recordings India. Note that visual cues probably also contribute to the differences in the perception.

The bar-plots of the responses are presented in Fig. 3. The left and right segments of the bar-plots present the statistics of the questionnaire responses from the recordings performed in Finland and India, respectively. Due to possible differences in the responses based on location and cultural background, we analyse the recordings from the two countries separately.

3.2.1. Finland

The experiences of privacy in the five scenarios are clearly different. For the primary question (Q1) in the office-closed scenario, the distribution follows our expectation that a majority of participants are comfortable to share private information. The office-open case presents a multi-modal distribution with peaks at *not at all likely* and *average* preference to share private information. The hall distribution has the lowest variance with most participants indicating least likelihood to share any private information. In the cafe-scenario, the mean response is neutral. The responses for outdoors suggest the people feel less private in such scenario. However, the high variance in the response might be a result of the difference in the traffic conditions for different recordings. While the responses to Q3 seem to be the inverse of those of Q1, we did not see a simple relationship between Q2 and Q1.

3.2.2. India

The distribution of the responses from India are clearly different from the distribution of responses from Finland. Both the office scenarios have similar trends, i.e. office-closed perceived highly private and office-open has a multi-modal distribution but with participants experiencing higher privacy in contrast to the responses from Finland. The responses for hall and outdoor scenarios are similar, despite the different characteristics of the environment in the two locations. The cafe-scenario has the distribution with the lowest variance, with most participants indicating a higher perception of privacy.

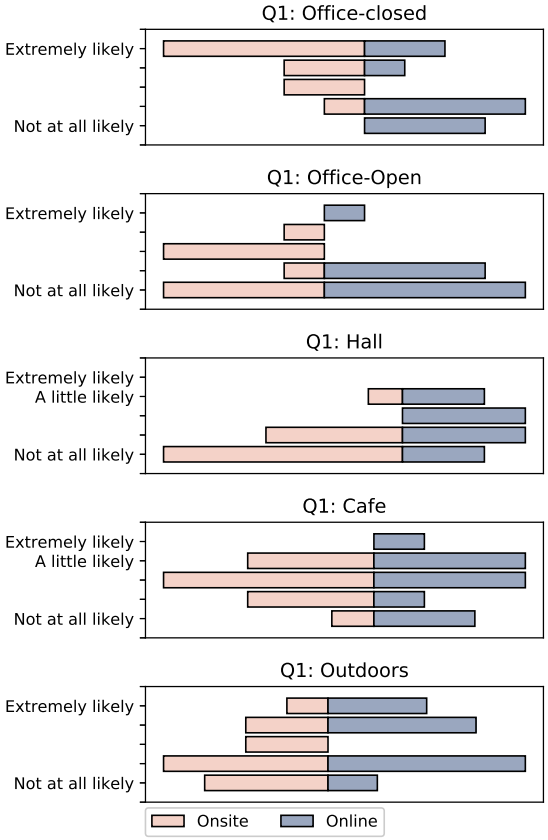

Figure 4: *Onsite versus Online responses to question Q1.*

4. Online questionnaire

Our aim is to quantify the relationship between the acoustic environment and the experience of privacy, and conversely, our hypothesis is that there is a relation between acoustic cues and the experience of privacy. Specifically, we assume that subjects can rate the experience of privacy similarly when they only hear the audio recording, as when they were present at the recording. To verify this claim and thus to validate our corpus, we designed an online experiment to rate different scenarios in terms of privacy from *only* acoustic cues, using the same questionnaire as in the onsite experiment.

The questionnaire was created using the browser-based listening-test platform *webMushra* [11], using the previously recorded audio samples. On each test page, participants listen to an audio sample representing one of the five acoustic environments and one of the three questions from Table 1.

We collected answers from 10 subjects. The purpose of the online questionnaire is to validate the corpus, hence this low number of subjects was deemed sufficient. We plan to amend this preliminary with a larger test in a later publication. Fig. 4 depicts the distribution of the responses to Q1, alongside the corresponding onsite responses from Finland for comparison.

We observe that both office scenarios present a low privacy rating. However, a minor percentage of participants perceive office-closed more private than office-open. For the hall-scenario, the average trend is around *average* and *not much likely*. The cafeteria is perceived to be the most private environment, which matches the hypotheses presented in Section 3.2, this could be due to highly non-stationary babble noise such that individual speech is incomprehensible. Additionally, the strong reverberation of the room could add to the noisiness of the scenario. The outdoors-scenario shows a prominent bi-modal distribution with means around *Not much likely* and *A little likely*. We leave a rigorous statistical analysis for later publications, but in Fig. 4 we see that responses in the onsite and online questionnaires have similar trends, thus validating our assumption that the audio signal carries cues which can be used to quantify the experience of privacy. Furthermore, some differences between the responses of the two questionnaires can be associated to the available visual cues during the onsite questionnaires.

5. Speech corpus structure

The database where the recordings are stored has been made publicly available at `http://soundprivacy.aalto.fi`. The database contains a directory with the audio files, license and the consent forms used in the recording process. The audio files are organised according to the location where they were recorded, the model of the recording device and the corresponding scenario.

The naming convention of the sound files is

$$scenario\_ID_1\_ID_2\_block.wav,$$

where ID_1 and ID_2 represent the ID numbers of respective speakers and *block* represents the corresponding 15 s segment in a recording. If only one subject interacts with one of the authors, the audio file is called *scenario_ID_block.wav*.

According to the general data protection regulations (GDPR) in the European Union [1], it is necessary to obtain the subjects' consent to store and process any of their personal data. A consent form was created to inform the users about the goals of the study and any possible use that their voices may have. The main points of our chosen level of privacy are:

- Data is anonymised such that the publicly available corpus contains only a unique ID number for each speaker.
- We secretly store a list linking subject names and their ID numbers for editing purposes.

The complete form can be found in the database repository.

6. Conclusion

We have recorded a speech corpus in real-world acoustic environments to quantify people's experience of privacy. Answers to a questionnaire complement the recordings and quantify the subjects' experience of privacy. To validate that this experience is encoded in the acoustic signal, the same questions were asked in an online questionnaire to subjects who had not been present during the recordings.

Our preliminary observations indicate that the acoustic information does affect the subjects' experience of privacy in different environments. In addition to auditory cues, other external aspects like visual cues will likely affect the perception of privacy in places with low levels of noise, where the acoustic information is limited. Internal factors like the personality or cultural aspects will also likely influence the results.

The correlation between the onsite and online results indicates that audio recordings of conversations in those environments carry information that people use to judge privacy. This matches our hypothesis and corroborates the validity of the presented corpus for the analysis of privacy. Thus, we expect that this corpus will provide researchers with useful data to analyse such effects on the experience of privacy.

7. References

[1] G. D. P. Regulation, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46," *Official Journal of the European Union (OJ)*, vol. 59, no. 1-88, p. 294, 2016.

[2] B. van der Sloot and A. de Groot, *The Handbook of Privacy Studies: an Interdisciplinary Introduction*. Amsterdam University Press, 2019.

[3] T. Bäckström, Ed., *Speech Coding with Code-Excited Linear Prediction*. Springer, 2017. [Online]. Available: http://www.springer.com/gp/book/9783319502021

[4] F. Dunn, W. Hartmann, D. Campbell, and N. H. Fletcher, *Springer handbook of acoustics*. Springer, 2015.

[5] M. Barron, *Auditorium acoustics and architectural design*. Routledge, 2009.

[6] S. S. Utami, J. Sarwono, N. Al Rochmadi, and N. Suheri, "Speech privacy and intelligibility in open-plan offices as an impact of sound-field diffuseness," in *Inter. noise*, vol. 2014, 2014, pp. 1–10.

[7] V. Hongisto, A. Haapakangas, H. Maula, and H. Koskela, "Simultaneous effect of office noise, heat, and stuffy air on employees' work performance," *Euronoise*, 2018.

[8] "IEEE recommended practice for speech quality measurements," *IEEE No 297-1969*, pp. 1–24, June 1969.

[9] "Audio recorder," downloaded: 10.10.2018. [Online]. Available: https://play.google.com/store/apps/details?id=com.sonymobile.\\androidapp.audiorecorder\&hl=en

[10] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[11] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA – a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.