Privacy and Smart Speakers

Guglielmo Maccario
Maurizio Naldi

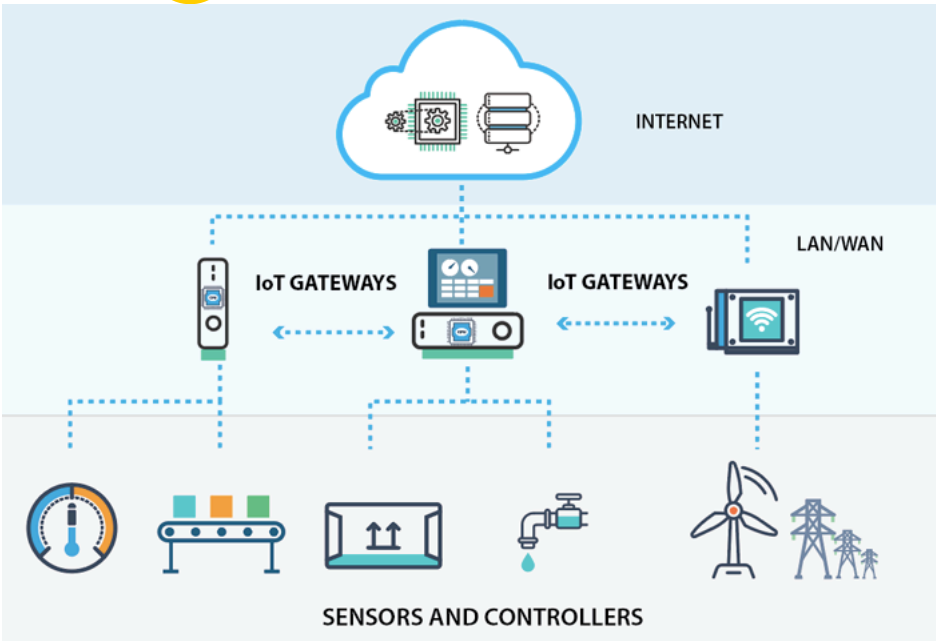
A brief background

Why Smart Speakers?

What is a Smart Speaker?

Smart speakers paired with intelligent virtual assistants like Alexa, Siri or Google Assistant, are a category of smart IoT devices that **take instructions** from users, **process** them, and **perform** the corresponding tasks, allowing users to carry out a wide range of tasks, completely **hands-free and eye-free**

Smart speakers and IoT



- From the IoT architecture viewpoint, a smart speaker is very similar to any other IoT device.
- It is equipped with sensors that measure physical properties of the environment. It is designed to record and send any voice interaction after hearing a wake-word.
- At the IoT cloud level, the data received is used to respond to the user request through virtual or physical actions and audio feedback

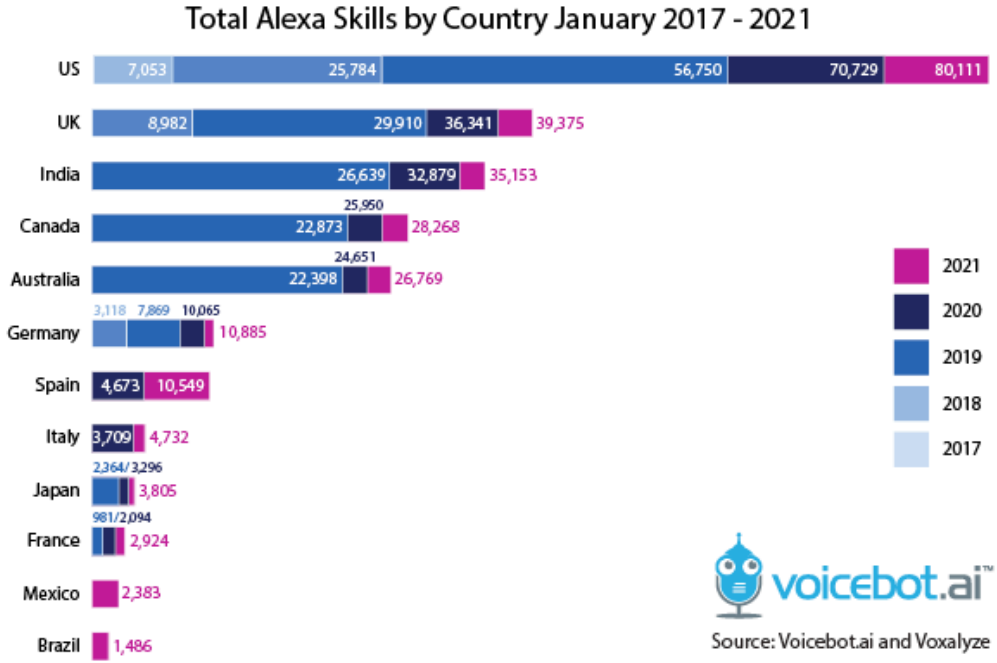
Beyond the definition...

Smart speakers are disrupting the consumer technology and service landscape.

Market value of Smart Speakers by 2025

35.500.000.000

US DOLLARS



- The number of tasks that a smart speaker can undertake, can be expanded with third party applications (”**skills**” and ”**actions**”)
- In 2021 Amazon has claimed over 100.000 Alexa skills worldwide.

Market Distribution

- Amazon is the leading vendor in the global smart speaker market, with a market share of 26.4% in Q3 2021.
- Google is Amazon’s closest competitor, with a share of 20.5% in the same quarter.
- Overall sales went up after Google's entry, from 6.57 million in 2016 to a projected 95.25 million in 2019.
- Chinese vendors Baidu, Alibaba and Xiaomi have become strong players thanks to growing demand in the Chinese domestic market.

However...

- Smart speakers are extremely **privacy-sensitive** devices: their recording and networking capabilities spur **concerns** about users' privacy.

- Smart speakers are vulnerable to **external attacks** that pose a threat for the privacy and security of their owners, **accidental triggering and misactivations**

- Outside the traditional academic publishing, the topic of smart speakers' privacy has gained attention also in **newspapers**

The Guardian
News website of the year
Search jobs
Sign in
Search
Sport
Culture
Lifestyle
More
Tech
Business
Obituaries
Environment
Science
Global development
Football
'Alexa, are you invading my privacy?' – the dark side of our voice assistants

StudyFinds
Research, in a nutshell.
Family
Health
Science
Society
Go ahead, I'm listening...

JAPANTODAY
Tech
Smart devices spy on you: How the Internet of Things can violate your privacy

Home » Google admits its Home speakers recorded at all times
Technologies
Management
Cyber
Security Newswire
Cyber Security News
Google admits its Home speakers recorded at all times

NEWS
RUSSIA-UKRAINE CONFLICT
POLITICS
ABORTION RIGHTS
PLAN YOUR VOTE
COVID
U.S. NEWS
OPINION
WATCH NOW
TECH & MEDIA
Alexa privacy fail highlights risks of smart speakers
Worries crop up about privacy issues surrounding personal info captured by Amazon's virtual assistant and competitors' similar connected speakers.

(© DedMityay - stock.adobe.com)
MOBILE PHONES, SCIENCE & TECHNOLOGY
Smart devices are watching you everywhere and violating your privacy, computer scientists warn

PARTNER CONTENT
IoT devices most vulnerable to internal security threats

Privacy is important = Users are concerned?

.

- Academics and practitioners have been discussing the privacy risks related to the use of smart speakers since the launch of the first Amazon Echo in the US market.
- Also, reports and surveys suggest that users are concerned for their own privacy and security
- What about **customer reviews** of smart speakers? Do they reflect the worries of academics and the general public?

Research questions

- **RQ1: Where in the world do we observe the highest interest in privacy issues in smart speakers?**
- **RQ2: Is there a trend for privacy issues interest?**
- **RQ3: What are the most investigated topics?**
- RQ4: **How do smart speaker users perceive privacy?**

RQ1: Where in the world do we observe the highest interest in privacy issues in smart speakers?

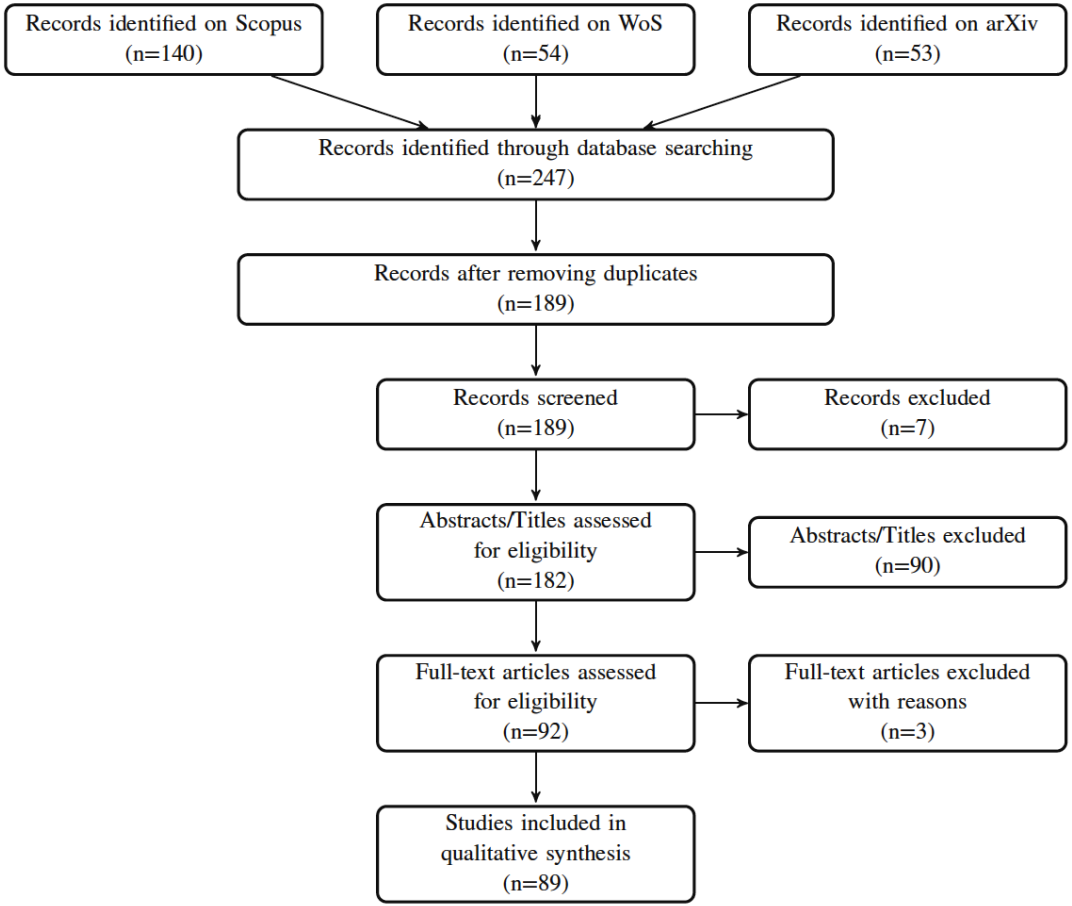
To answer RQ1 we performed a systematic literature review: the objective is to synthetize and analyse concepts, empirical findings, and issues related to the research on privacy in smart speakers

RQ1 Search Process and Eligibility

- **Subject**: we are interested in privacy issues related to the use of smart speakers. Since those devices have also been known as virtual assistants for a while, we have employed the following search string in our query: ("smart speakers" OR "virtual assistants") AND privacy.

- **Database**: we searched for papers on three major databases: Scopus, Web of Science (WoS), and arXiv.

Systematic selection flowchart



Two-level analysis

- **Meta-analysis** of the papers concerning publication outlets, time trends and geographical distribution of authors

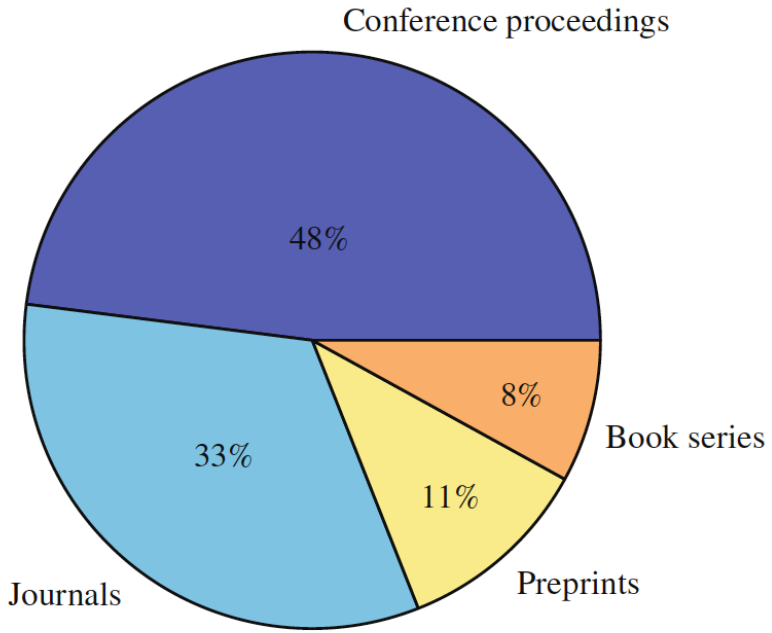
- **Topic analysis by** conducting a keyword analysis on the abstracts of the papers identified.

RQ2: Is there a trend for privacy issues interest?

To answer RQ2 we performed a meta analysis on: Pubblications outlets, time trends, geographical distribution

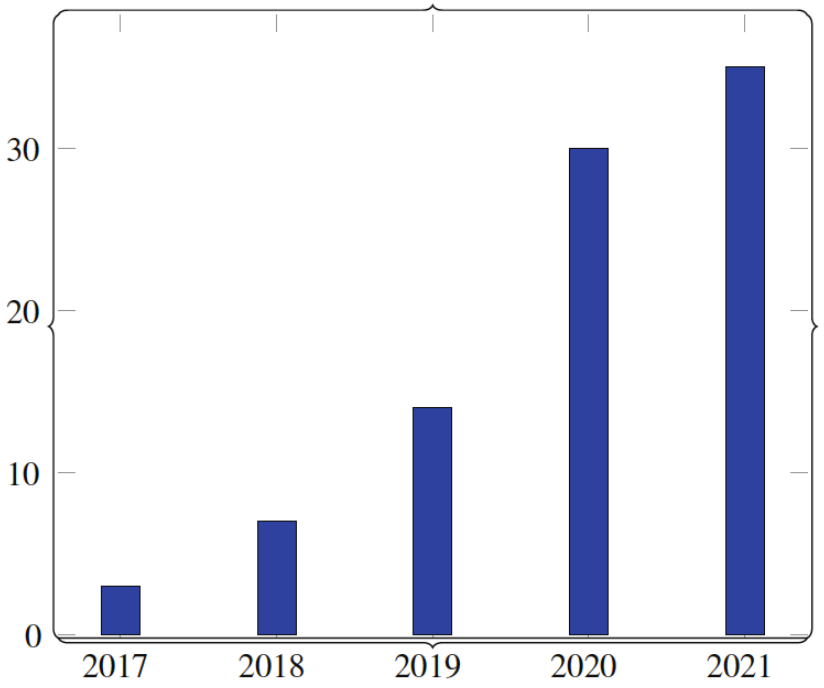
Meta-analysis: Publications outlets

- Dominance of the two traditional outlets: journals and conference proceedings (**together they account for 81%)**
- We also observe a significant fragmentation since **only two journals have published more than one article:** the International Journal of Human-Computer Interaction and Surveillance and Society



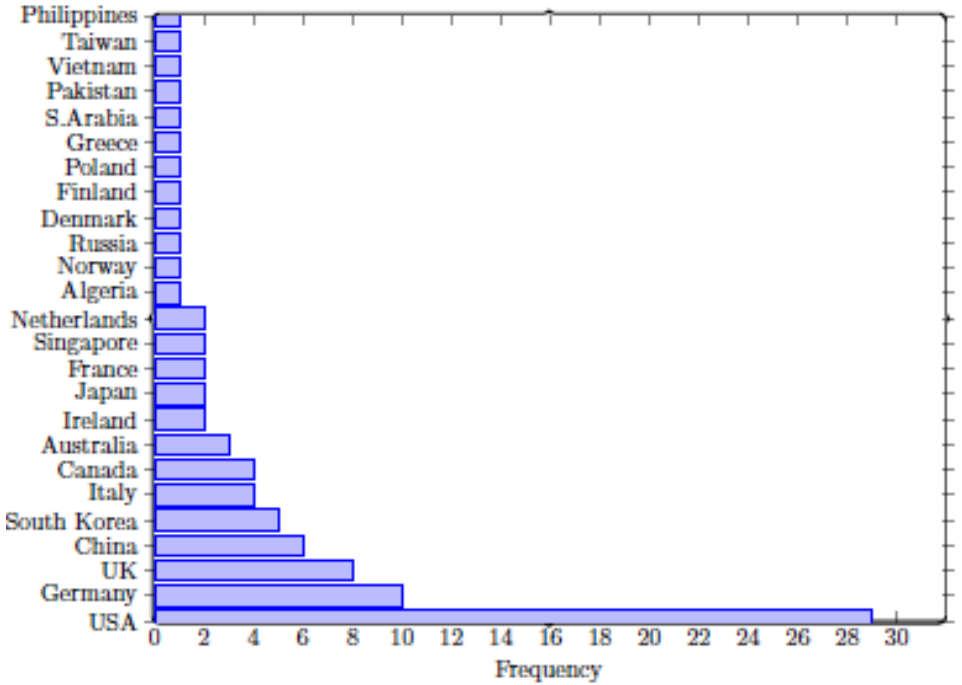
Meta-analysis: Time trends

- No literature on the topic before 2017. In that year, the first three papers on the subject appeared
- A continuous growth afterwards. The trend during the first four year with a nearly perfect doubling rate each year.
- The growth has continued in 2021 but at a reduced rate. The overall CAGR (Compound Annual Growth Rate) over 2017-2021 was anyway a large 63.45%.

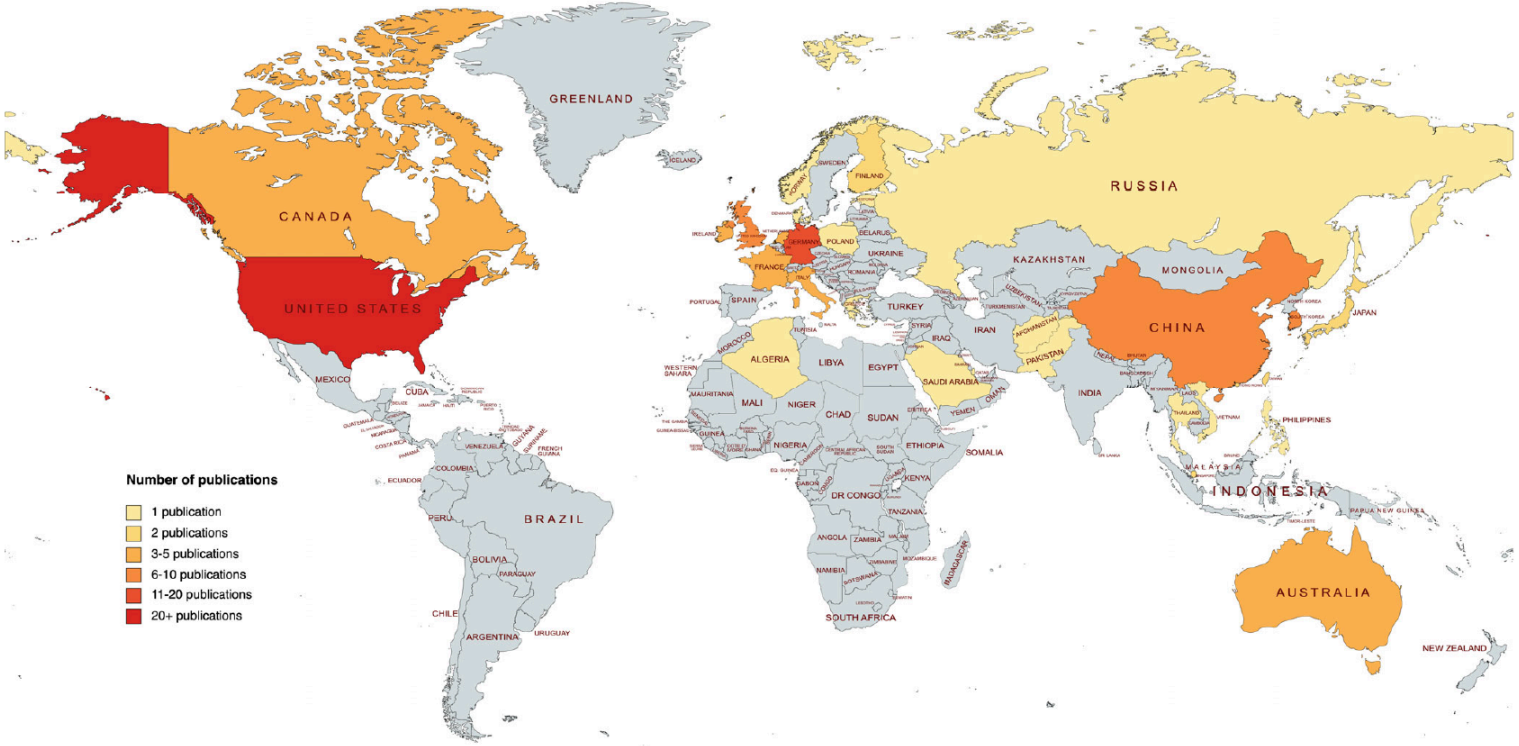


Meta-analysis: Geographical distribution

- Researchers are predominantly affiliated with institutes and universities based in the US
- US and China lead smart speakers ownership and production
- Since the first smart speaker was launched in the US in the 2015, the US market appears to be mature and represents a suitable context to study users' perception and privacy concerns



Meta-analysis: Geographical distribution



RQ3: What are the most investigated topics

Keyword and topic analysis

What are the most investigated topics?

- To answer RQ1 and RQ2, we have considered the papers' meta data, examining literature features that are not explicitly related to the contents.

- **To answer RQ3**, we turn to what the papers really talk about when dealing with privacy issues in smart speakers.

- We start by performing a **keyword analysis and then propose classifying the literature into essential topics.**

Keyword analysis

- We carried out an analysis to support the classification of the literature into essential topics.

- We conduct keyword analysis by extracting those keywords from the abstracts rather than relying on the keywords supplied by the authors.

- The shortlist of keywords provided by the authors may not reflect the actual contents of the paper, and the limitations usually imposed on the number of may result anyway in too brief a description of the paper's contents.

Keyword extraction

To extract keywords from the abstract we followed these steps:

1. Retrieving the abstract of each paper
2. Removing stopwords and other non-descriptive terms
3. Standardizing keywords
4. Sorting keywords in order of descending frequency

Top 20 most frequent words

Word	Frequency	Word	Frequency
Concern	80	Show	36
Information	55	Consumer	34
IoT	55	Behavior	30
Risk	52	Virtual	30
Interact	47	Design	29
Traffic	46	Learn	29
Model	44	Attack	28
Perceive	43	Experience	27
Personal	40	Accuracy	23
Recording	38	Older	22

First group of keywords

- Keywords neutral with respect to privacy perceptions and describing the object of analysis (not conveying any polarity, hence of little interest)
- Their overall presence in the top 20 is 37.2%.

Word	Frequency	Word	Frequency
Concern	80	Show	36
Information	55	Consumer	34
IoT	55	Behavior	30
Risk	52	Virtual	30
Interact	47	Design	29
Traffic	46	Learn	29
Model	44	Attack	28
Perceive	43	Experience	27
Personal	40	Accuracy	23
Recording	38	Older	22

Second group of keywords

- Words describing a relationship with smart speakers users.
- The presence of concern and risk show the negative perception of privacy-related issues as potentially dangerous ones
- This group is the most frequent one, accounting for 40.2%

Word	Frequency	Word	Frequency
Concern	80	Show	36
Information	55	Consumer	34
IoT	55	Behavior	30
Risk	52	Virtual	30
Interact	47	Design	29
Traffic	46	Learn	29
Model	44	Attack	28
Perceive	43	Experience	27
Personal	40	Accuracy	23
Recording	38	Older	22

Third group of keywords

- Actions carried out by/with smart speaker (some intrinsic to smart speakers' functions, such as interact)
- However, the most numerous subgroup refers to words having (potentially) negative connotation for privacy and security
- Though this group is the least numerous, it still accounts for 22.6%.

Word	Frequency	Word	Frequency
Concern	80	Show	36
Information	55	Consumer	34
IoT	55	Behavior	30
Risk	52	Virtual	30
Interact	47	Design	29
Traffic	46	Learn	29
Model	44	Attack	28
Perceive	43	Experience	27
Personal	40	Accuracy	23
Recording	38	Older	22

Topic analysis

- We try to classify papers according to their main topic (the specific aspect of privacy that the paper deals with).
- We perform a manual examination of the general theme of the paper rather than a bottom-up approach starting from the single keywords.
- We have identified five classes

Topics

- Privacy concerns
- Adoption factors
- Vulnerabilities
- Countermeasures
- Legal issues

Topic: Privacy concerns

The papers in this group examine the concerns of users and prospective users about privacy (sometimes verging on security issues).

The papers cover the whole trajectory that starts with:

- factors driving privacy concerns
- level of perception of privacy-related problems
- the impact of those concerns on users' behaviour.

Topic: Adoption factors (negative)

In the second class, we have papers that investigate the reasons and barriers to adoption.

Researchers seem to agree that the most recurring barrier is: **Presence of privacy concerns.** Some considerations on users:

- may be affected by bias (they have somewhat overcome their privacy concerns).
- appear willing to trade privacy for convenience.
- Do not use privacy controls of their devices
- Have reduced privacy concerns and enhanced trust if given the power to customize privacy settings

Topic: Adoption factors (positive)

Differences emerge instead as to the (positive) driving factors for adoption. Scholars found that these factor positively impact adoption:

- Platform-related variables, such as service availability, network size and complementarity with other devices
- The capability to connect to other household devices
- Technology optimism plus the capability to perform several functions
- Usefulness (and ease of use)
- Utilitarian and hedonic factors

Topic: Vulnerabilities

These papers examine how privacy may be compromised. Vulnerabilities can be classified into four classes:

- wiretapping,
- compromised devices enabled by smart speakers,
- malicious voice commands (fake or twisted)
- unintentional voice recordings

Topic: Vulnerabilities

Two peculiar attacks are described:

- Voice squatting: the attacker uses a word with a pronunciation similar to a legitimate word to invoke a malicious application
- Voice masquerading: a malicious application impersonates a legitimate application to steal their personal information.

Topic: Countermeasures

Papers in this group focus on how to prevent privacy and security leaks. Despite scholars seem to agree that most users seem to ignore the built-in countermeasures, some solutions are offered:

- Authentication to avoid unwanted users from issuing commands to the smart speaker
- Interposing a filter between the device and the server to avoid sensitive speech making its way to the server
- Avoiding recognition of commands through traffic analysis when the commands issued by the user find their way out of the local network to the server,

Topic: Legal Issues

Legal issues are present in a small but relevant number of papers.

Papers on this topic focus on:

- Lack of regulation regarding the sector
- Legal implications of recording guests without their explicit consent. Therefore, legislation against the manufacturers of speech-activated devices could be invoked.

RQ4: How do smart speaker users perceive privacy?

Text-minining and sentiment analysis in UK and South Korea

Text-mining analysis in UK

- To find out what consumers actually think about privacy in smart speakers, we performed a text-mining analysis of customers' perception of privacy in smart speakers.
- We analysed 4756 smart speakers product reviews posted on Amazon
- **The most owned smart speaker in the UK is Amazon Echo.**
- We collected reviews concerning three products of the Amazon Echo line: **Amazon Echo, Echo Dot, and Echo studio**.
- We examined the reviews after **pre-processing and filtering.**

Review Pre-processing

Pre-processing is needed to remove all the text that is not useful for later analysis and normalize the text format.

- Case Folding
- Stopwords filtering
- Removal of numbers, white spaces and punctuation
- Stemming
- Tokenisation

Review Filtering

To extract those reviews concernig privacy from the whole corpus of scraped reviews, we **need to assess the presence of privacy-related keywords.**

The construction of the set of privacy-related keywords required three parallel processes:

1. Identification of privacy synonyms
2. Identification of semantically similar words
3. Identification of co-appearing words in the literature

Review Filtering

- For the first stage, we formed a list of synonyms of the word privacy" by using the **Wordnet package**.
- However, out of 4756 reviews **only three** reviews contained any of those words.

```
> synonyms("privacy", "NOUN")
[1] "concealment" "privacy"     "privateness" "seclusion"   "secrecy"
```

Expansion of keyword list

- To expand the list, we conducted the second process: a survey among computer science students and obtained **eight additional privacy-related keywords**
- To **validate** these words and decide if they were a good fit for our research, we computed their **semantic similarity** to privacy.
- For that purpose, we employed Word2Vec.

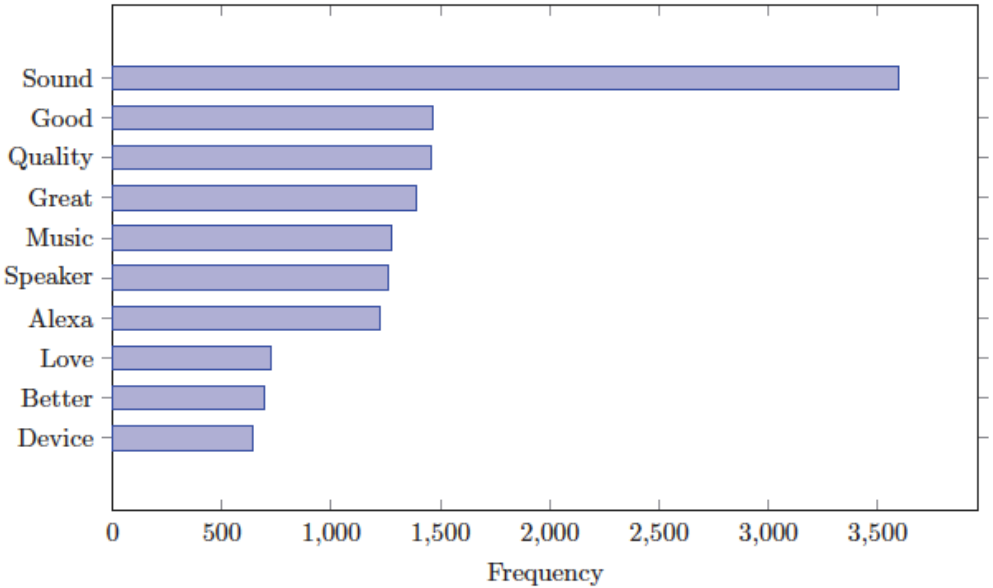
Word	Frequency
data	6
safe	6
trust	6
secure	4
privacy	3
intrusive	2
confidential	1
private	1
shield	1
Total	30

Table 4.3: privacy-related wordlist

- Word2Vec is an NLP technique to represent words as vectors
- We use cosine similarity to measure words semantic closeness
- Cosine similarity returns a similarity value ranging from –1 to 1,
- We used a corpus based on word embeddings trained on 6.5 billion words
- The similarity calculation returned good results for all 28 couples of words.

word1 \ word2	data	safe	trust	secure	privacy	confidential	private	shield
data	1.00	0.44	0.38	0.58	0.53	0.46	0.43	0.37
safe	0.44	1.00	0.76	0.71	0.72	0.69	0.73	0.57
trust	0.38	0.76	1.00	0.63	0.68	0.69	0.63	0.54
secure	0.58	0.71	0.63	1.00	0.80	0.62	0.62	0.52
privacy	0.53	0.72	0.68	0.80	1.00	0.64	0.62	0.50
confidential	0.46	0.69	0.69	0.62	0.64	1.00	0.66	0.47
private	0.43	0.73	0.63	0.62	0.62	0.66	1.00	0.51
shield	0.37	0.57	0.54	0.52	0.50	0.47	0.51	1.00

Table 4.2: Similarity matrix



- Even with the expanded keyword list, **only 6 reviews contained at least a privacy-related word**
- No privacy-related words appear in the top 10 most frequent words
- People appear to be concerned with audio and sound quality
- We then identified of co-appearing words in the literature.

- We examined the abstracts of the 71 papers obtained in the SLRc
- All the words (excepting stopwords) were extracted and sorted by frequency, retaining the top 50 words
- All the reviews containing any of the words in the top50 were provisionally added to the selection of useful reviews.
- Reviews were then manually inspected to assess if privacy-related.

Out of the **4,756** reviews making out our corpus, just **133** deal with privacy. Those reviews represent a meager **2.7%** of the whole corpus

Sentiment Analysis

We used two packages available for R, chosen because they perform well with short social media texts, do not require any training data, and are designed to identify negators:

- **SentimentR**
- **VADER.**

Sentiment Analysis

We found that:

- **The majority of customers who mention privacy-related words exhibit a positive sentiment.**
- **Despite privacy is of interest in a tiny number of reviews, it is a negative issue in half of them.**

Privacy concerns of smart speaker users in South Korea

- One limitation of our study was to be limited to UK
- In fact, the field of privacy of smart speakers has been largely limited to certain themes and national boundaries, lacking a cross-cultural, multinational approach.
- Therefore, we decided to examine the attitude of smart speakers owners towards privacy also in South Korea, replicating the methodology employed in the UK to compare a Western country and an Asian one.

Privacy concerns of smart speaker users in South Korea

We analysed 9500+ blog posts and product reviews, and we found that:

- South Korean smart speaker users seem to be either unconcerned or unaware of privacy issues.
- Like in the UK, even those few reviews that mention the topic were not found to carry a negative sentiment unequivocally.
- In comparison with our UK study, this investigation confirms the very limited interest in privacy issues, but shows an even lower degree of interest in an Asian country with respect to a Western one.

RQs Answers and Conclusions

RQ1 Where in the world do we observe the highest interest in privacy issues in smart speakers?

In the US: authors' affiliations (roughly 35%) couples with the strong presence of US manufacturers.

RQ2 Is there a trend for privacy issues interest?

YES. we observe a steadily growing trend. Privacy issues will become even more relevant when smart speakers reach a significant fraction of the population

RQ3 What are the most investigated topics?

The largest classes are countermeasures and vulnerabilities

RQ4 How do smart speaker users perceive privacy?

Privacy is not an issue for the overwhelming majority of smart speakers buyers

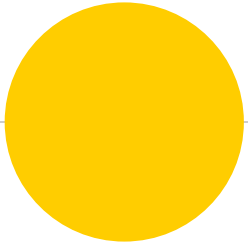
Publications

- Maccario, G., & Naldi, M. (2022). Alexa, Is My Data Safe? The (Ir) relevance of Privacy in Smart Speakers Reviews. *International Journal of Human–Computer Interaction*, 1-13.

- Maccario, G., & Naldi, M. Privacy in smart speakers: A systematic literature review. *Security and Privacy*, e274

Under review:

- Hong Joo Lee, Guglielmo Maccario, Maurizio Naldi; Privacy concerns of smart speaker users in South Korea: a text-mining analysis ; Journal of Open Innovation: Technology, Market, and Complexity



Thank you