
MAKING VOICES HEARD
NO PARKING
POST
ICE
Meow
BUY ONE GET ONE

Making Voices Heard

About the Study

We believe that voice interfaces have the potential to democratise the use of the internet by addressing limitations related to reading and writing on digital text-only platforms and devices. This report examines the current landscape of voice interfaces in India, with a focus on concerns related to privacy and data protection, linguistic barriers, and accessibility for persons with disabilities (PwDs). This project was undertaken with support by the Mozilla Corporation.

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Team

Research **SHWETA MOHANDAS, SAUMYAA NAIDU, DEEPIKA NS, DIVYA PINHEIRO, SWETA BISHT**

Conceptualisation, Planning, and Research Inputs **SUMANDRO CHATTAPADHYAY, PUTHIYA PURAYIL SNEHA**

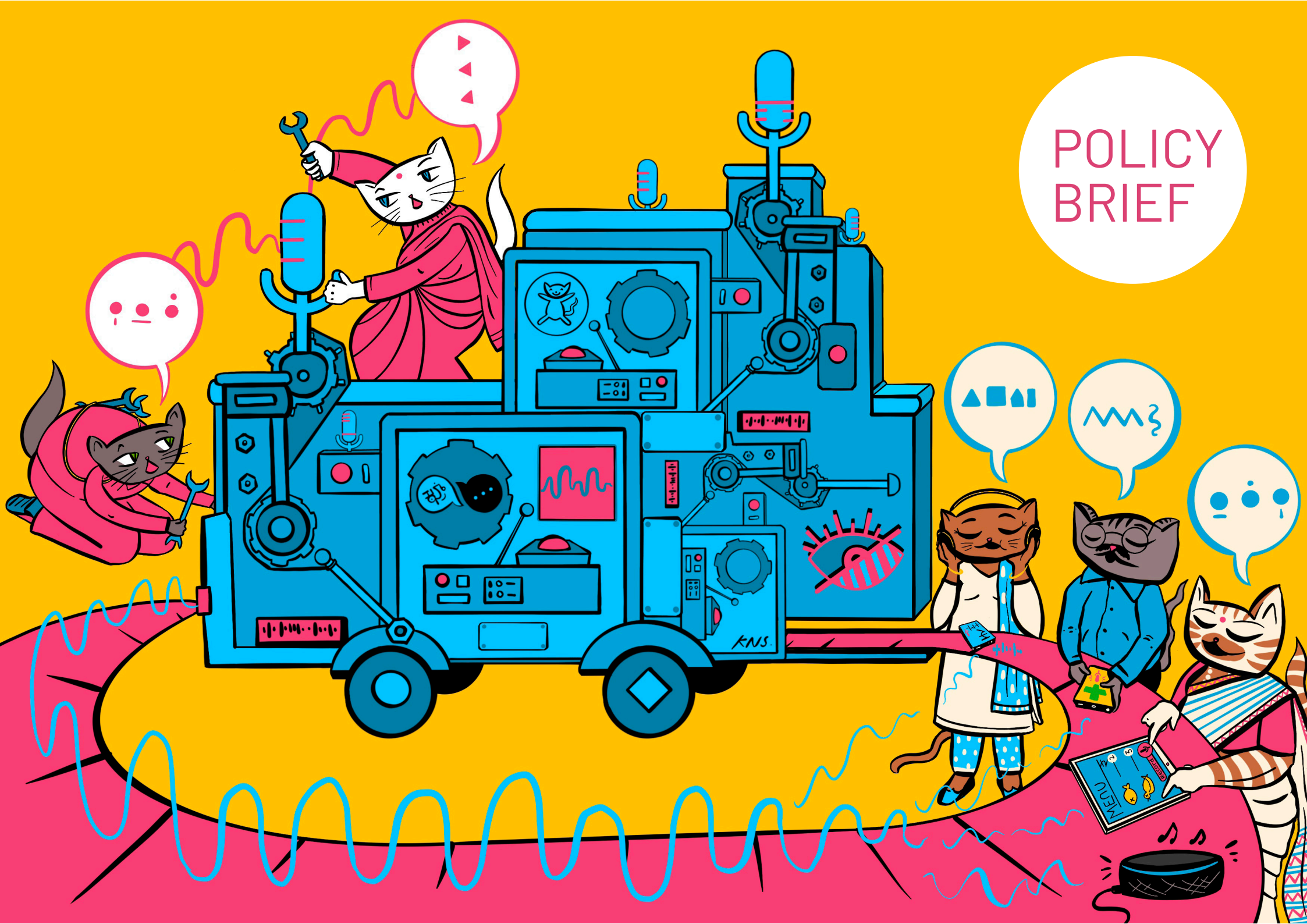
Illustrations **KRUTHIKA NS**

Report Layout and Design **SAUMYAA NAIDU**

Review and Editing **PUTHIYA PURAYIL SNEHA, DIVYANK KATIRA, PRANAV M BIDARE, TORSHA SARKAR, PALLAVI BEDI, DIVYA PINHEIRO**

Copy Editing **THE CLEAN COPY**

THE CENTRE FOR internet & society


POLICY
BRIEF
KNS
MENU

Making Voices Heard: Policy Brief

Research and Writing **SHWETA MOHANDAS**
Research Assistance **DIVYA PINHERO**
Review and Editing **PUTHIYA PURAYIL SNEHA, TORSHA SARKAR**
Research Inputs **SUMANDRO CHATTAPADHYAY**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. Introduction

Voice interfaces do not just provide an alternative way of interacting with a device; for people with low or no vision, they are the only way they can access the device. They allow people who are limited by text-only interfaces to navigate various aspects of their lives, by being able to access various services through voice. The development of voice technology has come a long way since the prototypes of the early 90s – they are now much cheaper, they can understand multiple languages and perform various tasks and can be integrated into different services. One of the earliest voice interfaces, the interactive voice response (IVR) system, emerged in the 1970s and is widely used even today. The technology has advanced by leaps and bounds since then, with the emergence of internet and smartphone-based voice interfaces that can be used to perform tasks of varying complexity, from setting alarms to ordering food.[1]

In India, given that IVR systems have been widely deployed for service delivery in both the public and private domains, there is a growing interest in internet-based voice interfaces that can understand multiple Indian languages. These interfaces have the potential to enable people to access services that were earlier restricted by language (English) and interface (text-based systems). Although there is vast potential, some of which has been harnessed by voice interface start-ups like Niki,[2] there is a need to ensure that these applications are available to people with varying accessibility needs. Given the current push towards more digital-first public services (e.g., the Cowin[3] platform), it is necessary to look at how accessible existing systems (such as websites) are and how voice interfaces can be integrated into them. Further, it is important to consider not just their potential but also the realities of a country where the infrastructural limitations can restrict access to services.

WIth respect to voice interfaces the advantages it can bring are curtailed by the unavailability of Indian language data. On the side of the individual there is also the need for better internet access to ensure that the people who will most benefit from voice interfaces can get to use them. Since voice interfaces are still in their beginning stages of uptake this is the right time to look at the challenges and possibilities towards their deployment. Additionally since the use of voice interfaces is still emerging in India, this is the right time to investigate the privacy concerns that may arise with the use of these interfaces and create policies in tandem with developments in data protection legislation.

This policy brief aims to bring into focus voice interfaces as an important policy question that needs more discussion and consideration, especially in India's quest for being a digital first nation. The policy brief also aims to shed light on the privacy concerns with respect to voice data, which seem to not get as much attention as facial data.

In light of these questions, this policy brief will look at the existing companies working on voice interfaces in India, the key concerns that limit their uptake, and the policy challenges in realising their potential.

1 Kozuch, K., "The 30 best Alexa skills in 2021", *Tom's Guide*, 4 August 2020, accessed 3 November 2021, https://www.tomsguide.com/round-up/best-alexa-skills.

2 "Niki." *Niki*. accessed 9 September 2021, http://niki.ai/

3 "CoWIN." *CoWIN*. accessed 9 September 2021, https://www.cowin.gov.in/

2. Voice interfaces in India

2.1. Mapping of actors in India

The voice interfaces ecosystem in India is slowly growing – a number of players provide voice services to businesses and consumers. However, when it comes to hardware-based voice interfaces, the key market players are Google[4] and Amazon,[5] which now support Indian languages spoken in different accents and integrate Indian apps (through Alexa skills) such as Ola.[6] To understand the state of voice interfaces in India, we mapped 27 voice interface developers in India, in terms of type of voice interface, client, sector, languages, and data collection. This revealed a few trends, based on the type of individuals they cater to, the sectors that use voice technologies extensively, and the most preferred languages, that could provide insights on the uptake of voice interfaces in the country.

More business-facing interfaces than consumer-facing

Although only Google and Amazon offer device-centric voice assistants,[7] a variety of mobile apps and smart devices incorporate voice interfaces. In our study of voice interfaces in India (including voice assistants), we were able to find only two apps – Niki[8] and Vokal[9] – that provided services to individuals directly. The remaining provided these services to businesses, which in turn offered them to the individual. Therefore, there are only a few general voice interfaces in India, as most are voice bots and chats developed for specific business purposes.

Sectors that use voice interfaces

The banking and finance sector features the highest number of chatbots and voice bots. These voice interfaces help individuals access information about their accounts as well as the services offered by the bank. HDFC Bank,[10] Andhra Bank,[11] and Kotak Bank[12] all use voice interfaces to interact with customers. The

4 Akolawala, T. "Amazon Echo Dot Tops Smart Speaker Sales in India in 2020, Google Home Mini, Mi Smart Speaker Follow: techARC." *Gadget360*, 18 February 2021, https://gadgets.ndtv.com/smart-home/news/amazon-echo-dot-most-sold-smart-speaker-india-2020-google-home-mini-mi-smart-speaker-techarc-report-2373059

5 Akolawala, "Amazon Echo", *Gadget360*,18 February 2021

6 "Ola", *Amazon*, https://www.amazon.in/ANI-Technologies-Pvt-Ltd-Ola/dp/B075NGT52M,18 February 2021

7 A program on a device that can listen and reply to voice commands.

8 "Niki." *Niki*.

9 "India's Largest Vernacular Question & Answers Platform in Indian Languages", *Vokal*, accessed 20 October 2021, https://www.vokal.in/

10 Ani, "HDFC's Banking CHATBOT 'Eva' Now Compatible with Google Assistant", *Business Standard*, 20 December 2017, accessed 20 October 2021 https://www.business-standard.com/article/news-ani/hdfc-s-banking-chatbot-eva-now-compatible-with-google-assistant-117122000272_1.html.

11 Hans News Service, "Andhra Bank Unveils AL Chatbot Abhi", *The Hans India*, 15 July 2019, accessed 20 October 2021, https://www.thehansindia.com/business/andhra-bank-unveils-al-chatbot-abhi-546877.

12 "Kota accessed 20 October 2021k Mahindra Bank Launches Keya – The First Voicebot in Indian Banking", *Kotak Mahindra*, 2 April 2018, https://www.kotak.com/content/dam/Kotak/about-us/media-press-releases/2018/kotak-mahindra-bank-launches-keya-the-first-voicebot-in-indian-banking-02042018.pdf.

second-most popular sector for voice interfaces is e-commerce, as apps such as Big Basket,[13] Grofers,[14] and Flipkart[15] use or have proposed to use voice interfaces. Some local governments also use voice interfaces services (offered through their websites or apps), such as the Rajkot Municipal Corporation and Pimpri Chinchwad smart city.

Languages

Hindi was the first and is still at times the only Indian language other than English available on virtual assistants and voice bots. Out of the 27 companies we mapped, all of them provided voice features in English and Hindi. Both Google Assistant[16] and Alexa[17] can understand and speak Hindi now. However Google and Amazon are yet to launch the voice assistant in other Indian languages. The other languages that follow Hindi in popularity are Tamil, Bengali, and Kannada.

Accessibility

Voice interfaces provide accessibility support for individuals who are unable to see the screen or understand the text. However, no applications other than Google and Amazon claim to provide accessibility features. Amazon Echo's website lists the various features that customers with vision, hearing, mobility, and speech accessibility needs could use.[18] Google Home provides accessibility features that allow the individual to control appliances and entertainment, make phone calls, broadcast messages, and manage tasks in addition to its voice assistant.[19]

Privacy

Voice interfaces have presented significant privacy concerns. The 'always on' feature of Google Home and Amazon Echo have attracted media attention for recording conversations even when the voice assistant was not summoned.[20] With respect to the voice interface companies that we analysed, it was difficult to assess privacy commitments as most developed voice interfaces for businesses, which then provided this service to customers. Hence, how these business-facing companies collect and store voice data is neither public nor addressed in their privacy policies. However, most companies developing voice

13 Rangarajan, K., "Voice to Cart: Powering your Ecommerce App with Voice", *Slang Labs*, 6 October 2020, accessed 20 October 2021, https://www.slanglabs.in/blog/voice-to-cart-powering-your-ecommerce-app-with-voice.

14 Limited, J. H. T., "How Haptik Automated Grofers' Customer Support in Less than 48 Hours", *Haptik*, accessed 20 October 2021, https://www.haptik.ai/resources/case-study/grofers-case-study.

15 Schwartz, E. H., "Indian E-commerce Giant Flipkart Expands English and HINDI Voice Search Platform-Wide", *Voicebot.ai*, 4 March 2021, https://voicebot.ai/2021/03/04/indian-e-commerce-giant-flipkart-expands-english-and-hindi-voice-search-platform-wide/.

16 Tech Desk, "Google Assistant Now in Hindi: Here's How to Activate and Use",. *The Indian Express*, 15 March. (2018, March 15). https://indianexpress.com/article/technology/social/google-assistant-now-available-in-hindi-heres-how-to-activate-and-use-5098595/.

17 Singh, M., (2019, September 18). "Amazon's Alexa Now Speaks Hindi",. *TechCrunch*., 18 September 2019, https://techcrunch.com/2019/09/18/amazon-alexa-hindi-india/.

18 "Accessibility Features for Alexa", *Amazon*, accessed 20 October 2021, https://www.amazon.in/gp/help/customer/display.html?nodeId=202158280.

19 "Accessibility features on Google nest or home devices", *Google Nest Help*, https://support.google.com/googlenest/answer/9286728?hl=en.

20 Guardian News and Media, "'Alexa, Are You Invading My Privacy?' – The Dark Side of our Voice Assistants", *The Guardian*, 9 October 2019, https://www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants.

interfaces have a publicly accessible privacy policy and terms and conditions. Some user-facing companies specified that they use, process, and store/retain voice data, whereas others failed to specify how they handle voice data. Although related laws, such as the Information Technology Act, 2001,[21] Sensitive Personal Data/Information Rules, 2011,[22] and Personal Data Protection Bill, 2019,[23] do not require companies to disclose if voice data is being processed, privacy policies that provide this information could help people make an informed choice of what they talk about or record on these applications.

3. Key concerns/questions

3.1. Questions around connectivity and infrastructure

The *Indian Telecom Services Performance Indicators* report published by the Telecom Regulatory Authority of India (TRAI) in 2020 revealed that as of 31 December 2019, there were 29.83 percent of rural internet subscribers in the country.[24] According to the license service area data that was provided, the states that had the lowest number of internet subscribers per 100 persons were Jammu and Kashmir (16 persons per 100) and Bihar and Uttar Pradesh (21 per persons per 100). The highest was Delhi (98.97 persons per 100).[25] The *Digital India* report of 2019 stated that India had 504 million active Internet users who were five years and above as of November 2019. In terms of usage frequency, nearly 70% of the internet-enabled population in India are daily users.[26] This data shows that although the number of internet users is large, the number of internet subscribers is still very low. This is due to the fact that in most households one smartphone is used by multiple people in the house.[27]

Thus, although several voice interfaces are being developed to cater to India's multilingual nature, they are limited in their reach until they can also be accessed by those without an internet connection or with intermittent access to the internet. A study on the use of IVR systems to support job searches by low-income domestic workers in India concluded that "for computer-based systems to solve developing-world problems often require significant work above and beyond an implementation of the technology."[28] Hence, although voice interfaces may benefit those limited by language and digital literacy, the proposed benefactors of the technology may be hindered by a lack of access to other key infrastructures.

21 The Information Technology Act, 2000.

22 Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011.

23 The Personal Data Protection Bill, 2019. http://164.100.47.4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf

24 Ministry of Communications, "Internet Connectivity in Rural India. Unstarred Question No. 594 To Be Answered On 16th September, 2020, 2020", 16 September, http://164.100.24.220/loksabhaquestions/annex/174/AU594.pdf

25 Ministry of Communications, "Internet Connectivity in Rural India".

26 Nandita Mathur, "India now has over 500 million active Internet users: IAMAI", *Mint*, 05 May 2020 https://www.livemint.com/news/india/india-now-has-over-500-million-active-internet-users-iamai-11588679804774.html

27 Dr Rajesh Tandon, "One Device Households", *The Times of India*, 17 July 2020. https://timesofindia.indiatimes.com/blogs/voices/one-device-households/

28 Smyth, T. N. (2010). Where There's a Will There's a Way: Mobile Media Sharing in Urban India. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. https://www.researchgate.net/publication/221514114_Where_there's_a_will_there's_a_way_Mobile_media_sharing_in_urban_india

3.2. The need for Indian language voice data

The developers and researchers interviewed for this study obtained voice training data from multiple sources such as open-source databases, at competitions set up by Google or Microsoft[29], user-generated anonymised data, databases like Mozilla's Common Voice, and hours of speech data recorded by professionals such as news readers or voice artists.

A common issue that the developers we interviewed highlighted was the scarcity of voice data in Indian languages. They noted that although there is now some data in Hindi and Indian English, there are several low-resource languages[30]. If data on them was available, voice interfaces could be developed to help people access services in these languages via their phones. The Indian scenario is particularly challenging due to the scarce availability of open-source voice data. Initiatives such as Indic TTS,[31] a consortium created and funded by the Government of India, have been making an effort to record data in various regional languages. However, finding the datasets and applying them to products is still a challenge. Another barrier that was highlighted was that technology giants such as Google and Amazon, with their abundant data and other resources, create an imbalance between start-ups that have to collect data from scratch and multinationals that already have data and systems in place.

3.3. Accessibility of government apps and websites

A 2012 study of 7,800 Indian government websites, which assessed their design against the Web Content Accessibility Guidelines (WCAG) 2.0, revealed that 1,985 websites failed to open and the remaining 5,815 had some form of accessibility barrier, including a lack of non-text alternatives to text making them inaccessible.[32] A more recent study, published in 2021, revealed that many government websites ranked low in usability, many did not follow WCAG 2.0 accessibility guidelines, and none of the 164 websites tested was fully accessible on mobile.[33] The study also stated that even in 2019, 62% of the websites they tested did not pass any MobileOK checks[34].

More recently, one of India's COVID-19 measures, the Arogya Setu app, and its mandatory use by citizens, have been debated strongly as it requires a phone and a working internet connection to access, apart from several concerns related to privacy and data protection. The app was also flagged by persons with visual or hearing disabilities and disability rights activists, for failing to meet accessibility standards. The

29 Through our interviews we understood that developers and researchers alike were able to get voice data in different languages through participating in competitions organised by Google and Microsoft.

30 A resource language means a language that does not have or has only few data resources. This makes it even more difficult to develop machine-learning based systems for these languages.

31 "Indic TTS", *Indic TTS*, **https://www.iitm.ac.in/donlab/tts/.accessed 3 November 2021**

32 "Accessibility of Government Websites in India: A Report", *The Centre for Internet and Society India*, 2012, **https://cis-india.org/accessibility/accessibility-of-government-websites-in-india**

33 Agrawal, G., Kumar, D., and Singh, M., "Assessing the Usability, Accessibility, and Mobile Readiness of E-government Websites: A Case Study in India", *Universal Access in the Information Society* (2021): 1–12.

34 The Mobile Ok checked by W3C performs various tests on a web page to determine the level of mobile-friendliness. The tests are defined in the mobileOK Basic Tests 1.0 specification. A web page is considered mobileOK only when it passes all the tests.

Union Social Justice and Empowerment Ministry informed the Ministry of Electronics and Information Technology (MeitY) and the National Informatics Centre (NIC), that the app lacked accessibility features.[35] A report by activist Anjlee Agarwal stated that the visually impaired people who tested the app found it inaccessible, which amounted to a violation of the Rights of Persons with Disabilities Act, 2016. According to the report, the "the screen reader in the app did not announce the purpose of all controls or the type of control, whether a link or button". This means that the screen reader did not specify what tasks or options the app could provide, and it did not differentiate between whether there was a link or a button to enter the service. The app also did not mention the page numbers on the website, which would mean that the individual might miss out on the next page or the screen reader would keep on reading the pages on a loop. Additionally, on the "Your status", "COVID updates", and "E-Pass" tabs in the app, "the screen reader was not announcing the control type, so individuals did not know these were interactive tabs."[36] In May 2020, an IVR service was set up within Arogya Setu to aid people who had feature phones and landlines.[37] However, there were no known improvements with respect to the accessibility of the Arogya Setu app itself.[38] The Supreme Court, while examining issues relating to COVID-19 management, emphasised the need to conduct a disability audit for the CoWIN website and Aarogya Setu to ensure that they were accessible.[39]

Hence, for India, there is a need not just for the implementation of voice interfaces, but also for other accessibility measures to be introduced to enable every person to benefit from the digital world.

3.4. Emerging uses of voice and questions about privacy and data protection

Despite their several benefits, particularly in terms of enabling individuals to access the internet and services in their own languages, voice interfaces present significant privacy concerns. Researchers and civil society have raised concerns regarding the potential for misuse and harm that might stem from storing and processing immense amounts of voice data. These recordings may have been made without the person's knowledge and may reveal extremely sensitive information – its most benign consequences range from targeted ads to being profiled based on what the device processes. One of the emerging concerns is how this voice data could be shared with law enforcement agencies and the consequences of such sharing.

Additionally, there seems to be a growing interest in using voice as a biometric identifier, especially in the banking sector. A report by Kaizen Secure Voiz detailed the benefits of voice biometrics such as fraud detection, rural banking, and remote verification.[40] However, the report also recognised

35 Nath, D., "Mandatory Aarogya Setu App Not Accessible to Persons with Disabilities", *The Hindu*, 2 May 2020, https://www.thehindu.com/news/national/coronavirus-mandatory-aarogya-setu-app-not-accessible-to-persons-with-disabilities/article31489933.ece.

36 Nath, D., "Mandatory Aarogya Setu", *The Hindu*.

37 "Arogya Setu IVRS", https://www.mohfw.gov.in/pdf/AAROGYASETUIVRS1921.pdf

38 Nath, D., "Mandatory Aarogya Setu", *The Hindu*.

39 "In Re: Distribution of Essential Supplies and Services During Pandemic", In The Supreme Court Of India Civil Original Jurisdiction, 2021, https://main.sci.gov.in/supremecourt/2021/11001/11001_2021_35_301_28040_Judgement_31-May-2021.pdf

40 Kulkarni, A., "Indian Banking – Adoption of Voice Biometrics", 2020, https://kaizenvoiz.com/wp-content/uploads/2020/11/Kaizen-white-paper-for-Indian-banking-ver-6.1.pdf

the challenges that would come with switching to voice biometrics, such as user confidence (making the person confident in using their voice, and confidence in the safety of using voice), training of staff and capacity of the organisation implementing it. Some banks that have looked at implementing voice recognition are Citi Bank, HSBC, and Standard Chartered Bank, which seem to have implemented this in India as well. However, implementation of voice biometrics should also come with adequately addressing the privacy and data protection responsibilities of collecting and processing biometric data (in this case, voice data).

4. Policy recommendations

4.1. The impetus for public-funded research

A project at the scale of Indic TTS was possible because of the availability of government funding. There is a need for increased public funding of voice-based research in Indian languages to allow researchers and developers to create localised voice interfaces. However, one of the issues with publicly funded research is that open access research and databases require continuous funding to be sustainable. Unlike private for-profit companies, public-funded research or datasets are usually made available free of cost.

In the case of Indic TTS, the datasets are all open access and can be used by start-ups and researchers alike; the objective is to allow more projects and research questions to stem from the existing work and to foster an environment of collaborative, open-access research. Our conversation with Indian start-ups working on voice revealed that they mainly relied on datasets from large companies such as Google for their voice data which these startups either purchased or won as a part of challenges organised by the companies. While initiatives such as Indic TTS do exist, there seems to be a disconnect between researchers and start-ups working on voice in Indian languages. One way to foster innovation is to have public–private partnerships that would not only ensure that the research is relevant to the needs of the industry but also that the industry benefits from the research and the development. Another way to boost further research on voice interfaces specifically for Indian languages could be to set up a system of royalty-free licensing for start-ups, where once the start-up starts to seek commercial value from the datasets, the license can be changed to a revenue-sharing model.[41] This system would ensure that the researchers receive feedback after deploying the research in the real world and the start-ups can test and verify the same. The above system could be beneficial for start-ups that do not have the capacity or the funding to set up public–private partnerships.

4.2. More funding for accessibility research

There has been a worrying decline in budgetary allocations towards schemes for persons with disabilities in India. The budget for the Scheme for Implementation of Persons with Disabilities Act (SIPDA) was cut from INR 315 crore in 2019–20 to INR 252 crore—a 20 percent reduction—in 2020–21. Similarly, the budgetary allocation for both research on disability-related technology and the National Institute of Mental Health and Rehabilitation in FY 2020–21 was missing, compared to INR 20 crore in the previous year.[42] The assistance for Disabled

41 Ali, F. and Mohandas, S., "The Compulsive Patent Hoarding Disorder", *The Hindu*, 24 March 2017, https://www.thehindu.com/opinion/op-ed/the-compulsive-patent-hoarding-disorder/article17617888.ece.

42 Ali, A, "Scheme for Implementation of Persons with Disabilities Act (SIPDA) Has Been Reduced from Rs 315 Crore", *Indian Express*, 30 January 2021, 20, https://indianexpress.com/article/lifestyle/life-style/pandemic-has-hit-persons-with-disabilities-hardest-union-budget-should-address-their-concerns-7167840/.

Persons for Purchase (ADIP)/Fitting of Aids and Appliances has also not seen any increase in allocation of funds and stands at INR 230 crore for the entire population of persons with disabilities.[43] The national pre-budget consultation held by the National Centre for Promotion of Employment for Disabled People (NCPEDP) emphasised the need to incentivise companies that make accessibility products (both hardware and ICT) by providing rebates and concessions.[44] As recently as August 2021, the Standing Committee on Social Justice and Empowerment (Department of Empowerment of Persons with Disabilities) expressed that the progress of the Accessible India Campaign, launched in 2015, has been "rather slow".[45] The campaign aims to make accessing services such as transport, public spaces, tourist places, international airports, railway stations, and information and communication technology in India easily accessible for persons with disabilities.

4.3. More clarity from Personal Data Protection Bill about the regulation of voice data

The Indian Personal Data Protection Bill, in its 2019 version, defines biometric data as "facial images, fingerprints, iris scans, or any other similar personal data resulting from measurements or technical processing operations carried out on physical, physiological, or behavioural characteristics of a data principal, which allow or confirm the unique identification of that natural person."[46] Although voice data has not been explicitly mentioned in this definition, it could fall under the processing of the physical characteristics of the data principal, which are unique to each individual. Biometric data is also considered sensitive personal data; hence, requirements such as the need for explicit consent to collect, share, store, and use such data, and the prohibition of processing such data outside India, are being established under the PDP Bill. The Bill also mentions an additional category of data fiduciaries called significant data fiduciaries,[47] which have more duties and responsibilities based on the volume of data processed, the sensitivity of that data, risk of harm, and the use of technologies. The Bill also states that if in the opinion of the Data Protection Authority, data processing by a fiduciary carries risk of significant harm to any data principal, then that fiduciary will be tasked with all or some of the responsibilities of a significant data fiduciary.[48]

Although voice data can be considered biometric data and is in the ambit of sensitive personal data, it needs to be clearly included in the definition of biometric data in the Personal Data Protection Bill. This is becoming increasingly crucial as several services are including voice data, and certain institutions, such as banks, have also begun to use voice biometrics as a form of recognition.[49] This would mean that a person's voice can be linked to their financial information, thus linking two types of sensitive information to a service or a company.

43 Ali, "Scheme for Implementation", *Indian Express*.

44 Ali, "Scheme for Implementation", *Indian Express*.

45 Outlook, "Progress of Accessible India Campaign Rather slow: Parl Panel." *Outlook*, 6 August 2021, **https://www.outlookindia.com/newsscroll/progress-of-accessible-india-campaign-rather-slow-parl-panel/2136476**.

46 Section 3(7), The Personal Data Protection Bill, 2019, **http://164.100.47.4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf**

47 Section 26(1), The Personal Data Protection Bill, 2019, **http://164.100.47.4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf**

48 Section 26(3), The Personal Data Protection Bill, 2019, **http://164.100.47.4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf**

49 "Making Voices Heard: Mapping Actors," *Making Voices Heard*, accessed 02 February 2022, **http://voice.cis-india.org/mapping-actors.html**

4.4. The need for more diverse voice datasets

Hindi was the first and is still the only Indian language available on some voice interfaces, both for virtual assistants and voice bots.[50] The mapping of voice interfaces in India revealed that out of the 27 companies covered, all provided voice features in English and Hindi. Hindi is also the Indian language of choice used in the most popular voice assistants, Amazon's Alexa and Google Home. One of the reasons why Hindi is used so widely in voice interfaces is because it is one of the few high-resource languages in India with multiple voice datasets. Private companies develop voice interfaces for the most popular or most spoken languages as they are more profitable. The creation of voice databases for lesser spoken languages is left to volunteer-based organisations and public-funded projects. There is a need to look at how voice interfaces can be encouraged to support more Indian languages. Although there are several IVR systems in different Indian languages, their scope is limited to particular questions and answers.

4.5. The need for more funding towards community-led voice dataset collection

When a handful of companies are made responsible for collecting, processing, and creating speech datasets, the choice of languages is based on popularity and commercial viability. Even these systems, which work with data-rich languages, often fail to understand accents and voice modulations that are not present in the datasets.[51] Additionally, as these datasets are owned by large corporations, they are protected by non-disclosure agreements, contracts, and Intellectual property rights. However, as stated by one of our interviewees, "language technology is an entry into a digital world",[52] especially in a country with widespread inequity in access to digital infrastructure. Community based voice data collection initiatives are attempting to bridge this gap by assembling open-access datasets.

In India, the Indic TTS consortium was created with the goal of making information available in regional languages. However, due to the scale and the resources required, the consortium could only collect data for 13 Indian languages. Common Voice[53] (a global open-access dataset of voice recordings in multiple languages that can be used to train speech-enabled applications) is another great example of how a community-driven and open-access collection of voice data can lead to a more inclusive internet. Common Voice now has over 13,905 hours of voice data across 76 different languages as of July 2021.[54] This was achieved by not only making the language available on Common Voice, but also by making the website available in that language. When adding a new language, the community localises 85% of the website, so that the local language community can easily navigate it without relying on

50 Ahaskar, A. "Voice biometrics are Cleverer Now, But Still Need More Work", *Mint*, 6 February 2020, https://www.livemint.com/technology/tech-news/voice-biometrics-are-cleverer-now-but-still-need-more-work-11581011267941.html.

51 WP Company. "The Accent GAP: How Amazon's and Google's smart SPEAKERS Leave Certain Voices Behind", *The Washington Post*, 19 July 2018, https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.

52 Interview, Anonymous, in person, Bangalore, March 3 2020.

53 "Making Voices Heard: Common Voice Case Study," *Making Voices Heard*, accessed 02 February 2022, http://voice.cis-india.org/case-studies/common-voice.html

54 "Common Voice by Mozilla." *Common Voice*, accessed January 4, 2022, https://commonvoice.mozilla.org/en/datasets.

English. When a language is active on the site, it is up to the community to present 5,000 sentences in that language that can be recorded. This indicates two things to Common Voice: one, that there is an active language community that can provide language recordings, and two, that the barrier to get the language into Common Voice is fairly low.[55]

A recent example of community-driven voice data collection initiative was for Kinyarwanda, a widely spoken language in Rwanda with over 12 million speakers.[56] In 2019, Mozilla and Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) co-hosted an ideation hackathon in Kigali to create a data corpus for Kinyarwanda. A result of the hackathon was Digital Umuganda, a volunteer-driven start-up with the aim to build digital infrastructure such as voice data. Despite the challenges faced in mobilising the community, including poor access to mobile phones and the prohibitive cost of data plans, the startup managed to collect 1, 211 hours of Kinyarwanda voice data from a diverse set of over 420 contributors.[57] They are planning to set up a hybrid model involving both on-site and off-site recording through in-person and online events and by mobilising an expanding pool of volunteers. They hope that this process will hasten the contributions and be capable of withstanding any unforeseen circumstances.[58] One of the ways India could look at increasing the language reach of voice interfaces is to learn from the example of Rwanda, and have initiatives that bring together government agencies, startups and student volunteers to create voice data in languages from each state and community. In India, the CGNet Swara project is a great example of how voice can be used to help individuals of a particular community. CG Net Swara[59] is an Indian voice-based online portal that serves as a platform to discuss issues related to the Central Gondwana region in India. People in the forested regions of Chhattisgarh use it to report and share news in the Gondhi language through a phone call. Gondhi, which is spoken by almost 2 million people in different parts of northern and western India, can only be written by 100 people.[60] This is where a voice-based interface for people to report stories and listen to them in Gondhi helps. The portal is accessible through mobile phone or desktop; people can also listen to news reports and stories by giving a missed call. The CGNet Swara website helps the community preserve their language by participating online and via phone.

For government initiatives and private players, studying the approaches and best practices adopted by projects such as Common Voice and CGNet Swara could help expand their work and thereby the reach of the internet. Initiatives and projects such as these help reduce the language barrier, improve access to infrastructure and public services, provide services to people

55 "Making Voices Heard: Common Voice Case Study," *Making Voices Heard*, accessed 02 February 2022, **http://voice.cis-india.org/case-studies/common-voice.html**

56 "How Rwanda is making voice tech more open", *Mozilla Foundation*, 16 September 2020, **https://foundation.mozilla.org/en/blog/how-rwanda-making-voice-tech-more-open/**.

57 "How Rwanda is", *Mozilla Foundation*.

58 "How Rwanda is", *Mozilla Foundation*.

59 "Welcome to CGNet Swara", *CG Net Swara*, **http://cgnetswara.org/**,16 September 2020

60 Majumdar, M., "This Indian Language Can Be Written by Only 100 People", *The Hindu*, 31 March 2018, **https://www.thehindu.com/society/this-indian-language-can-be-written-by-only-100-people/article23384526.ece**.

across languages and digital literacy, help people learn new skills, enhance adherence to privacy and accessibility guidelines, and help preserve low-resource and indigenous languages.

5. Conclusion

Voice interfaces have immense potential to make the internet accessible to people who are limited by purely text-based interfaces. However, in the case of India, there needs to be greater research and policy discussions on the challenges, possibilities, and dangers of voice interfaces. Currently, the discourse around voice interfaces has been sporadic, with announcements that certain government services will be accessible through voice but without much follow-up.[61] There is also a need to look at how public and private services can be made universally accessible to people with varying accessibility needs. Additionally, accessibility should not be the sole responsibility of the government; private companies and start-ups should assess how accessible their services are, conduct user research, and have people with various accessibility needs on their teams. Along with the possibilities that voice interfaces bring, there is also a need to consider the privacy concerns and potential harm that they can cause. Given the possibility of widespread use of voice biometrics, there is a need to ensure that voice data is not used for profiling. Voice data should be given the same significance as facial recognition data, and how such technology is being deployed should be examined.

To sum up, voice interfaces and voice data have immense potential in India; however, greater attention needs to be given to development of policies directly related to these technologies. This would ensure that their full potential is reached without harming the individual using it or creating language erosion.

61 For example there have been numerous news reports about the Umang App being enabled with multilingual voice support, however at the time of writing this policy brief there have been no reports of its implementation and use.

Appendix – Timeline of key voice interface events

Government initiatives

Owing to the language diversity and low literacy rate of India, a number of studies and initiatives have studied IVR systems, including Avaj Otalo[62] (a service for farmers to access relevant and timely agricultural information) and Sehat ki Vaani[63] (for the management of Type 2 diabetes and maternal health). In the year 2020 the Aarogya Setu IVRS service was set up to check the spread of COVID-19 and help people detect symptoms.[64]

Although there have been no policies yet that directly regulate and encourage the uptake of voice interfaces, a few government initiatives encourage the development and adoption of voice technologies. One such initiative is the Indic TTS platform, sponsored by DeiTY, Ministry of Information Technology. The goal of this initiative is to develop a corpus of text-to-speech data in Indian languages. The consortium includes some of India's premier institutions, and the researchers have been able to collect a total of 40 hours of speech data in 13 Indian languages so far.

Umang

In 2018, the Indian government announced the inclusion of a multilingual voice search feature in the Unified Mobile Application for New-age Governance (UMANG) platform. Developed by the Ministry of Electronics and Information Technology and National e-Governance Division, UMANG provides easy access to an array of government services via smartphones and on their website.

Although the UMANG website and app are currently not enabled with voice technology, government tenders published in February 2020 reveal that the government intends to create a conversational chatbot and AI-based voice assistant.[65] They also emphasised the need to include more Indian languages to ensure inclusivity and widespread adoption. More recently, in 2021, the Ministry of Electronics and IT selected Senseforth AI as the firm to provide these services on the Umang platform. The first deployment will include voice bots and chatbots in English and Hindi, after which the service will expand to Malayalam, Tamil, and Telugu.[66]

'AIRAWAT' (AI Research, Analytics, and Knowledge Assimilation platform)

In January 2020, NITI Aayog released an approach paper to set

62 "Voice-based Social Media", *Awaaz.De*, https://hci.stanford.edu/research/voice4all/, 16 September 2020

63 Kazakos, K., Asthana, S., Balaam, M., Duggal, M., Holden, A., Jamir, L., Kannuri, N. K., Kumar, S., Manindla, A. R., Manikam, S. A., Murthy, G. V. S., Nahar, P., Phillimore, P., Sathyanath, S., Singh, P., Singh, M., Wright, P., Yadav, D., and Olivier, P., "A Real-time IVR Platform for Community Radio", proceedings of the 2016 CHI Conference on Human Factors in Computing System, 2016. https://doi.org/10.1145/2858036.2858585

64 "Arogya Setu IVRS", https://www.mohfw.gov.in/pdf/AAROGYASETUIVRS1921.pdf

65 "Invitation to Bid for Appointment of Partner Agency (Vendor 5)", Umang, https://www.meity.gov.in/writereaddata/files/tender_upload/UMANG%20RFP_AI-Bot.pdf

66 Agarwal, Surabhi, "Move Over Alexa and Siri, 'Hey Umang' to Deliver Govt Services Through Voice Commands Soon", *Economic Times*, 05 April 2021, https://economictimes.indiatimes.com/tech/technology/move-over-alexa-and-siri-hey-umang-to-deliver-govt-services-through-voice-commands-soon/articleshow/81916003.cms

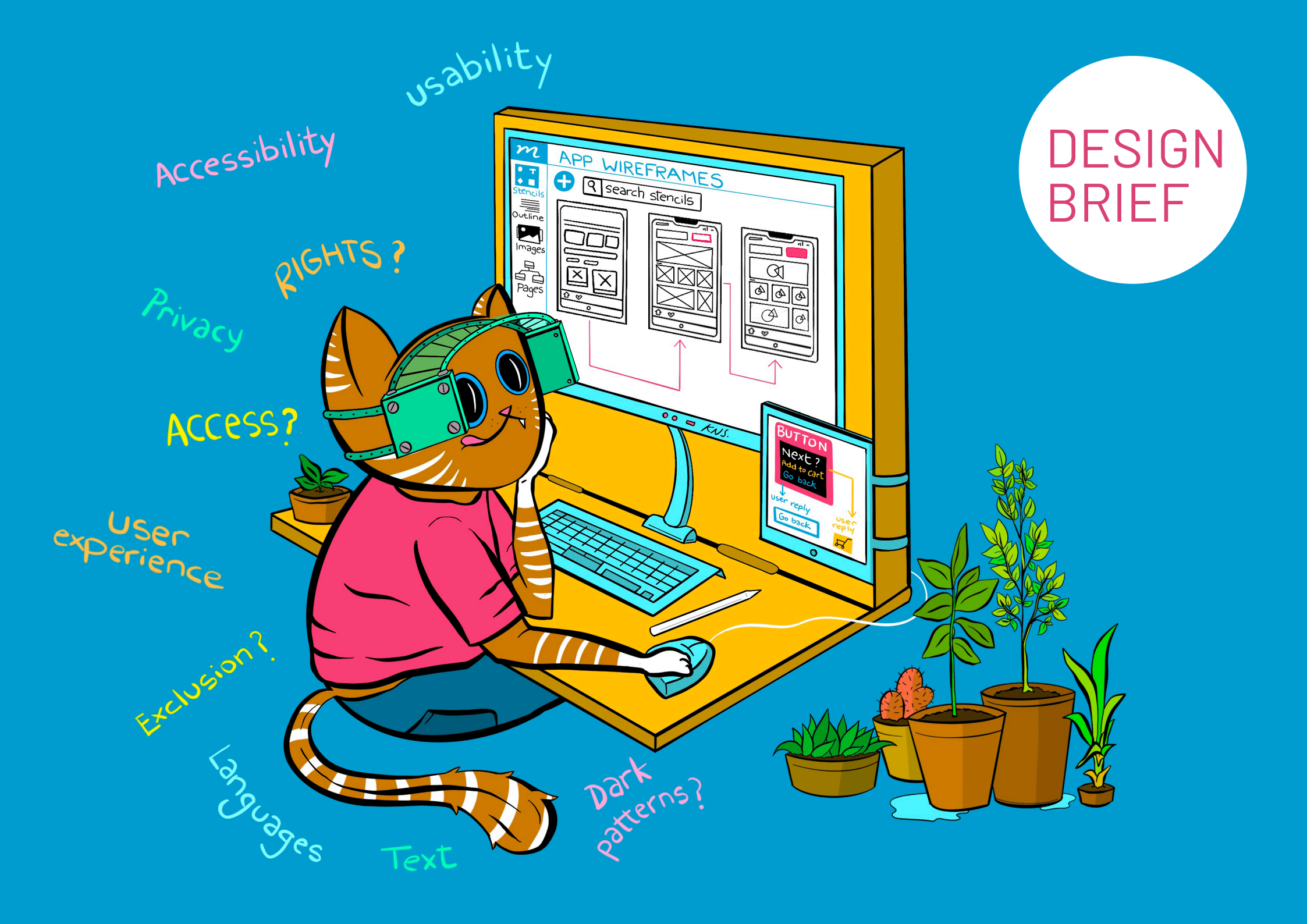
up India's first AI-specific cloud computing infrastructure, called 'AIRAWAT' (AI Research, Analytics and Knowledge Assimilation) platform. In the AI strategy paper released in 2018, Niti Aayog stated that the cloud-based platform would support AI-based speech recognition and natural language processing for research and development.

State initiatives

The Tamil Nadu government, under the Tamil Nadu e-governance agency (TNeGA), has expressed interest in creating a voice user interface in Tamil for availing of government services. Santosh Mishra, the chief executive officer of the Tamil Nadu e-Governance Agency (TNeGA), also stated at the summit on Responsible Artificial Intelligence for Social Empowerment (RAISE) that the voice interface would ensure that "the keyboard barrier to access technology is lifted".[67] With respect to existing voice services, the Madurai Kavalan app is a good example – the app allows individuals to record voice-based police complaints. The user study revealed that the voice API helped older people and those who found it hard to type and navigate the menu to access the app. The emergency feature also provides a 'women's safety' option, where a woman can either press the emergency button or request for help by saying "help me" in English or Tamil, which would trigger an SOS response.

The Bangalore Electricity Supply Company Ltd (BESCOM) has reportedly been working with the Machine and Language Learning (MALL) Lab at the Indian Institute of Sciences (IISc) to develop an "artificial intelligence-powered voice bot to attend to customer calls". This voice bot is being designed to allow people to seek answers to basic queries in English and Kannada.

67 Shivakumar, C., "TN Agency to Develop First Voice User Interface by Government in Tamil", *New Indian Express,* 9 October 2020, https://www.newindianexpress.com/states/tamil-nadu/2020/oct/09/tn-agency-to-develop-first-voice-user-interface-by-government-in-tamil-2208051.html.


DESIGN BRIEF
usability
Accessibility
RIGHTS ?
Privacy
ACCESS?
User experience
Exclusion?
Languages
Text
Dark patterns?
APP WIREFRAMES
search stencils
Stencils
Outline
Images
Pages
BUTTON
Next ?
Add to cart
Go back
user reply
Go back
user reply

Making Voices Heard: Design Brief

Research and Writing **SAUMYAA NAIDU**
Research Assistance **SWETA BISHT, DEEPIKA NANDAGUDI SRINIVASA**
Review and Editing **SHWETA MOHANDAS, PUTHIYA PURAYIL SNEHA, DIVYANK KATIRA**
Research Inputs **SUMANDRO CHATTAPADHYAY**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. Background

Given the increasing number of voice interface (VI) products in India, it is important to understand and analyse their design as part of the research and development of these technologies. In order to understand the challenges and opportunities in VI design, as well as to identify some best practices, we interviewed designers working with VIs.

The existing VI landscape comprises various actors, such as start-ups, global organisations, developers, policy-makers, and individuals using the VIs. Our mapping of actors in the VI industry suggests that a large number of the upcoming VI products in India are being developed by private companies.[1] In these companies, VI products are mostly conceptualised by developers, and designed by in-house teams. These teams include designers specialising in conversational design, and user interface (UI) and user experience (UX) design. Conversational experience design is still an emerging discipline in the country. It draws from human conversation patterns to make digital systems easy and intuitive to use.[2] The principles of conversational design can, therefore, be applied beyond voice assistants and chatbots to include all UI and web design.[3]

However, in the case of most start-ups, the designers' role and scope are based on the UI and UX design process.

Our primary methodology involved interviews with designers working independently or with developers, start-ups, and global organisations, as well as developers who make broad design decisions and work with designers. These designers and developers include Preeti Sheokand, a user experience designer who specialises in conversational design at Symphony AI;[4] Kumar Rangarajan and Vinayak Jhunjhunwala, co-founder and marketing associate, respectively, at Slang Labs;[5] Jai Nanavati, co-founder at Navana Tech;[6] Megan Branson, senior product designer at NVIDIA, formerly at Common Voice;[7] Akshay Kore, senior product designer at Observe.ai;[8] and Keshav Prawasi, co-founder at Niki.[9]

In this study, we look at the design of VIs in India based on three key criteria: multi-language support, accessibility, and privacy. As VIs gain popularity in the country, developers have realised that multi-language support is the key to success with Indian audiences. The focus on multi-language support has been a business enabler for the VI landscape. It has opened up a new design space, with a shift in focus from metros to start-

1 "Making Voices Heard: Mapping Actors," *Making Voices Heard*, accessed 02 February 2022, http://voice.cis-india.org/mapping-actors.html

2 Hampton, M., "Principles of Conversational Design," *Marvel Blog*, 30 October 2020, accessed 4 August 2021, https://marvelapp.com/blog/principles-of-conversational-design/#:~:text=The%20concept%20of%20conversational%20design,more%20natural%20dialogue%20with%20systems.

3 Hampton, "Principles of Conversational Design."

4 "Symphonyai: Transforming Businesses with Enterprise AI," *Symphony AI*, accessed 4 August 4 2021, from https://www.symphonyai.com/.

5 "Slang Labs: Add Accurate Multilingual Voice Assistants to Your App," *Slang Labs*, accessed 4 August 2021, https://slanglabs.in/.

6 "Navana Tech: Turn on the Conversation," *Navana Tech*, accessed 4 August 2021, https://navanatech.in/.

7 "Common Voice by Mozilla," *Common Voice*, accessed 4 August 2021, https://commonvoice.mozilla.org/en.

8 "Contact Center AI," *Contact Center AI | Observe.AI*, accessed 4 August 2021, https://www.observe.ai/.

9 "Niki: Aapke Ghar Ki Manager," *Niki*, accessed 4 August 4 2021, http://niki.ai/.

ups aiming at the 'next billion users'.[10,11,12] However, developers have also identified challenges in supporting the numerous languages and dialects in India such as lack of language training data and technical expertise. Further, though VIs are globally recognised as accessibility tools for people with disabilities, at present, accessibility is not considered a primary objective for VI products in India. Instead, VIs are seen as tools to reach low-literacy individuals. In terms of privacy, the concerns around VIs usually focus on the device always listening for the 'wake word'.[13] There are no comprehensive guidelines on privacy standards for VIs. The design of VIs hence, also do not follow any specific privacy-preserving principles.

2. VI design and development processes

Over the course of our conversations with designers and companies, we observed that there is a largely standard process for the design and development of VIs. Based on the design thinking process, it comprises the broad steps of primary research, conversation-design modelling, testing, and refining.

2.1. Primary research

Preeti Sheokand explained that her design process begins with secondary research on the domain and the context within which the VI product will operate. She then identifies the limitations or challenges present in the context – for example, the presence of a noisy environment. Further, she conducts primary research to understand the needs and context of the people who are going to use the VI. Based on this, she works out the intents of the VI, including its 'hygiene intents'. An intent is the objective of the voice interaction or the individual's intention.[14] The VI understands these intents and responds to them. 'Hygiene intents', as Preeti calls them, are interactions needed to accomplish the intent. For example, if the objective is to buy groceries, the hygiene intent would include signing up or signing into the platform, creating a profile, and selecting items for purchase.

Preeti observed that while technologists create synthetic conversations for development purposes, there is a lack of understanding of natural conversations. She addresses this using the training data collected during the primary research. She then extrapolates starting points based on this research. She then works on the conversation flows following UX and conversation-experience design principles. Preeti explained that conversational design, when using artificial intelligence (AI), is a probabilistic system and not a deterministic one. In

10 Majumdar, S., "Voice and Vernacular: The Future of E-retail in India," *Fortune India: Business News, Strategy, Finance and Corporate Insight*, 27 February 2021, https://www.fortuneindia.com/first-edit/voice-and-vernacular-the-future-of-e-retail-in-india/104630.

11 Choudhury, D., "Building Products for the Next Billion Users: Solving the Language Barrier, Monetisation Puzzle and More," *Inc42 Media*, 26 September 2020. https://inc42.com/features/decoding-the-psychology-of-the-next-billion-users-as-products-scale-up/.

12 Sachitanand, R., "Voice, Video and Vernacular: India's Internet Landscape is Changing to Tap New Users," *The Economic Times*, 7 October 2018, https://economictimes.indiatimes.com/tech/internet/voice-video-and-vernacular-indias-internet-landscape-is-changing-to-tap-next-wave-of-users/articleshow/66102478.cms?from=mdr.

13 Lynskey, D., "'Alexa, Are You Invading My Privacy?' – The Dark Side of Our Voice Assistants,," *The Guardian*, 9 October 2019, https://www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants.

14 "What Is a Voice User Interface (VUI)?" *Alan Blog*, accessed 20 May 2021, https://alan.app/blog/voiceuserinterface/.

a probabilistic system, the occurrence of events cannot be predicted perfectly.[15] The behaviour of such a system can be understood in terms of probability. A deterministic system, on the other hand, is one in which the occurrence of all events is known with certainty.[16] The VI may have multiple responses based on what the individual says and how the VI understands it. The various possibilities of how a VI system understands a phrase are determined by its technological limitations and the individual's context. Hence, the conversation flows are decided for each of these possibilities or responses.

Our conversations with Navana Tech and Niki indicated that VIs are currently being envisioned for audiences with less experience with technology, and, hence, the initial design process revolves around capturing their interactions with existing platforms and assessing how they would potentially interact with VIs. Navana Tech has divided its audience into five literacy and technology cohorts by conducting tests to determine each segment.

The team at Niki too has significantly invested in primary research. Over the last three years, the team has spent about 30,000 hours talking to individuals using the Niki app. Keshav Prawasi informed us that Niki has a dedicated in-house customer insights and research team that consistently works towards understanding these individuals better. Their team of researchers, designers, and product managers has also travelled to Rajasthan and visited Tier 2, Tier 3, and other cities, like Chomu, Pushkar, Ajmer, and Udaipur, to conduct usability studies. They studied how people interact with platforms such as YouTube and WhatsApp. Further, they conducted hackathons, brainstormed for a few weeks, and recorded videos with people. They ideated multiple design concepts and finalised a few, based on which they built prototypes. They then tested specific use cases such as bill payment. They studied people's interaction with the prototypes and further refined the design. Then they ran tests again to observe for which functions people relied more on voice. Finally, they made upgrades and changes to reflect these observations.

Kumar Rangarajan at Slang Labs described the user research that he conducted for product design with Srishti Manipal Institute of Art, Design, and Technology, Bengaluru. The design researchers worked with some apps that used touch and others that used speech. They spoke to more than 50 people, including some who are not well-versed with technology but who owned at least a smartphone. They asked shopkeepers and people on the streets and at bus stands how they would communicate to have a device perform a certain task. They also conducted more detailed, one to two-hour-long interviews with 10 people and observed them. They carried out primary research in English and Hindi. Then, they identified a use case and designed wireframes to test the flow of the interface. Like Preeti, Kumar also talked about considering the various intentions of the individuals using VIs, creating all possible conversation flows, and identifying different variables in these flows. He also spoke of expanding design details by adding various ways in which a primary use case, such as voice search, can be triggered while using the VI. Their focus is on 'productising' all the learnings from the research into their VI platform, so that businesses who integrate the VI do not need to start over.

15 Thakur, D., "Differentiate between Deterministic and Probabilistic Systems," *Computer Notes*, 30 January 2013, accessed 22 May 2021, **https://ecomputernotes.com/mis/information-and-system-concepts/differentiate-between-deterministic-and-probabilistic-systems**.

16 Thakur, D., "Differentiate between Deterministic and Probabilistic Systems."

2.2. Understanding the context

Going deep into the design process, Akshay Kore, who has previously been part of the team working on the Microsoft Cortana interface, shared several insights. To begin, he shared some questions to consider while designing a VI. The first is whether a voice interface is appropriate to that particular context. To answer this, Akshay shared some advantages of VIs and contexts where it is suitable:

- The VI substitutes for a complex action or task that requires multiple clicks or steps. For example, setting an alarm requires multiple steps and is time-consuming, but through VI, it can be set using a single command.
- When people are engaged in an activity where they cannot use their hands to access technology, voice becomes an important medium to interact with the device (for example, while driving or cooking).
- Using VI does not require any added learning. People can ask the VI questions, and the VI can either answer the question or respond that it cannot understand or address the query.
- Talking is more intuitive than other ways of interacting with technology. One can convey more through a VI than through a text-only interface, as other factors, such as tone, the difference between a question and an exclamation, and several emotions are conveyed more effectively through voice.

Akshay also mentioned contexts that are inappropriate for VIs:

- In public spaces, it is difficult to use VIs due to the presence of noise.
- If an application requires a lot of editing, using a VI is not advisable, as one cannot undo actions on VIs.
- Individuals may not be comfortable sharing health-related information or other private details with a machine, especially if the VI is being used in a public space.

Akshay reiterated Preeti's idea of context – he suggested that the designer be aware of the context for which they are designing the VI. They should consider the surroundings of the individual, whether they are a beginner or an expert in using technology, and the type of device they are using (a device with a screen or a speaker or both). For example, when designing a healthcare-based VI, research may not be easily available as it comprises sensitive information, so the design process would need to involve several rounds of testing and feedback.

2.3. Testing and refining

Megan Branson, former senior product designer at Common Voice (CV), mentioned that they applied design at a conceptual level. The project used an iterative process which involved repeated testing with people and refining the platform. The project began with identifying the need for large quantities of publicly available voice data that could be used to train speech-to-text engines. Design thinking exercises with Mozilla community members were conducted to ideate on creating an open-source voice dataset.[17] Megan created paper prototypes of design concepts and gathered feedback on them. The initial assumption was that people would need an ulterior motive to share voice data. However, their research revealed that most people were willing to donate voice data. The team also realised that people wanted to learn more about the need for voice data collection. Hence, they designed a platform whose predominant

17 Branson, M., "We're Intentionally Designing Open Experiences, Here's Why," *Medium*, accessed 13 May 2021, **https://medium.com/mozilla-open-innovation/were-intentionally-designing-open-experiences-here-s-why-c6ae9730de54**.

objective is collecting voice data.[18]

In this initial iteration, CV developed an interactive model where people could 'teach' a robot to understand human speech by reading sentences to it.[19] This version intended "to tell the story of voice data and how it relates to the need for diversity and inclusivity in speech technology".[20] The team then gathered community feedback and developed further iterations. Megan explained that they did a UX audit of the working prototype at this stage and made further refinements. Since 2017, they have focused on improving the platform – primarily improving the experience of contributing voice data. They also took UX heuristics, competitor evaluations, and community feedback into consideration.

Our interviews indicate that there is a strong emphasis on primary research to understand the needs of people from varying backgrounds. Most interviewees placed a lot of focus on understanding how people with less experience of technology, in both rural and urban settings, use VI. Many VI companies aim to provide reliable banking and fintech services for rural audiences. As there is little precedent for designing VIs in India, designers follow the established UI/UX path. Most designers working on VI products are UI/UX or product designers by training or experience and have only recently familiarised themselves with the nuances of conversational experience design.

2.4. Conversation experience design

Conversation design is only just emerging as a discipline and specialised practice in India. Designers and services have put together guidelines and principles of conversational design.[21] The core principles for conversation design in India were developed by Cathy Pearl in her book *Designing Voice User Interfaces*.[22] During his presentation on designing the best VI experiences, Vinayak Jhunjhunwala from Slang Labs talked about the best practices for conversation experience design based on Pearl's book and other resources.

- Defining expectations using convention: In VI design, it is important to break away from existing conventions and unlearn previous digital behaviour.
- Setting the right expectations: It is important to eliminate open-ended greetings and rhetorical questions from the design, as they are difficult to answer for the VI due to cognitive overload.
- Discoverability: Elements should be easily accessible in the VI. For example, the individual should be able to discover the voice button and quickly understand how to use it.
- Affordance: Vinayak emphasised that voice needs novel affordance strategies, such as audio prompting and adding visual depth to the interface.
- Fail-safes: Fail-safes should be built into the interface to counter instances of when a phrase is not heard, or when it

18 Branson, M., "We're Intentionally Designing Open Experiences."

19 Branson, M., "We're Intentionally Designing Open Experiences."

20 Branson, M., "We're Intentionally Designing Open Experiences."

21 "Voice Principles: Clearleft," *Voice Principles | Clearleft*, accessed 7 June 7 2021, **https://www.voiceprinciples.com/**.

22 Pearl, C., *Designing Voice User Interfaces: Principles of Conversational Experiences*, O'Reilly Media, (2017).

is heard incorrectly.

- Use cases: He recommended picking common use cases, creating sample dialogues for each case, and testing them with different people. Sketching a voice user interface (VUI) flow diagram is another recommended technique.
- Confirmations: There should be explicit audio confirmation of commands to assure the individual that the VI has understood the task.
- Error Handling: VI needs to be designed to handle errors and latency in responses.

Other principles in Pearl's book include using conversational markers that let the individual know where they are in the conversation; adapting to the experience and expertise of novice and expert individuals keeping track of the context of the input; including a set of universals such as 'repeat', 'main menu', and 'help', at every stage; using audible or visual cues to communicate unavoidable system delays; designing experiences for accessibility; and prioritising personalisation over personality.[23]

3. Challenges in designing VUI

Some of the key challenges that designers faced were the poor memory of the VI; the need for multiple potential conversation flows; the need to handle errors; technological barriers; and a lack of language compatibility.

3.1. Poor memory

Akshay warned us that maintaining a record of previous interactions is a concern for VIs. While designing, it is important to ascertain what sort of memory the machine should have. This can be difficult to judge, as in some cases, multiple individuals may access the device. Products such as Amazon Echo are designed for the home setting, where there are multiple family members and activities.[24] But most voice assistants are designed to talk to one person at a time. The design challenge here is creating voice assistants that can address a group, know how many people are present, and be able to distinguish the situations and profiles of these individuals.[25] Akshay suggested that profiling the individuals talking to the device can enable it to provide contextual responses based on who is interacting with it. While it is not clear what this profiling would entail, it could mean differentiating individuals based on their speech patterns and/or voice biometrics.[26] This can then be used to build a history of their commands and identify and list their intent. However, voice profiling can have grave privacy implications, such as voice-based surveillance, targeted advertising, and the leakage of sensitive voice biometrics.[27]

3.2. Designing potential dialogues

Akshay also stated that the information provided by the VI must be related to the conversational context. While humans understand this intuitively, automated responses may not

23 "Voice Principles," Voice Principles.

24 Santos, M. E., "Designing Better Voice interfaces for Everyday Life," *Medium*, accessed 23 June 2021, https://uxdesign.cc/designing-better-voice-interfaces-for-everyday-life-2cb344913fae.

25 Santos, M. E., "Designing Better Voice Interfaces."

26 Turow, J., "Shhhh, They're Listening – Inside the Coming Voice-profiling Revolution," *The Conversation*, 28 April 2021. https://theconversation.com/shhhh-theyre-listening-inside-the-coming-voice-profiling-revolution-158921.

27 Turow, J., "Shhhh, They're Listening."

always be appropriate. Niki also refers to this challenge as "bringing the individual back into the conversation". Thus, it is necessary to write rigorously and design potential dialogues between the interface and the individual. As we mentioned earlier, VI uses probabilistic technology, which means there can be multiple responses in different contexts. For instance, how a VI interprets homophones such as 'pair' or 'pear' would depend on the context. In the case of a food delivery app, 'pear' takes precedence over 'pair'. These potential dialogues must be designed to understand the questions accurately, respond appropriately, set the right expectation, and provide confirmation to the individual.

3.3. Handling errors

Many variables affect VIs – such as background noise and accents – and which make them less than perfect. Handling errors becomes significant when designing VIs. The accuracy of the device, or the confidence of output, as Akshay phrases it, is impacted by design and product thinking.

3.4. Focus on technical approaches

Preeti pointed out that the VI industry has only recently realised the value of UX. Most developers are not designing for experiences but for the completion of tasks. The interfaces are mostly created by people with technical expertise. Preeti suggested that it is critical, even from a business perspective, that one moves beyond this purely technical approach and starts looking at how individuals use VI. The focus should shift from data sets to studying use cases; the design process requires greater sensitivity towards the purpose and the audience and their comfort.

3.5. Controlling or restricting content

Preeti also emphasised the need to focus on accessibility, transparency, and ethical practice when designing VIs, as the applications are a lot more open-ended. To illustrate her point, she used the example of Alexa, where children may ask the device for information that their parents may not want them to know yet. While parental controls can be applied to visual content, controlling or restricting content on VIs is more challenging as identifying and differentiating between the individuals using the device is a complex operation.

3.6. Designing with optionality

Navana Tech's co-founder, Jai Nanavati, also talked about the challenge of dealing with optionality in voice menus. He explained that although banking facilities can offer 20–30 service options, it is difficult for the individual to remember the options and effectively provide input to the VI. He suggests that a chat-based interface is more helpful in such a scenario. Effective VUIs should provide brief information and ask individuals if they want to hear more before offering additional options.[28] It is also important to allow individuals to request that information is repeated whenever they need it.[29]

3.7. Clarifying scope

Kumar observed that while touch limits options to those on the screen, in the case of VI, the individual using it can say anything.

28 Sengupta, A., "A Sound Relationship: 4 Tips to Build an Engaging Voice User Interface," *Wipro Digital*, accessed 18 May 2021, https://wiprodigital.com/2019/05/22/a-sound-relationship-4-tips-to-build-an-engaging-voice-user-interface/.

29 Kamm, C., "User Interfaces for Voice Applications," in *Voice Communication between Humans and Machines*, National Academies Press: OpenBook, 2019, 426 https://www.nap.edu/read/2308/chapter/30#426.

Hence, when an individual says something that is beyond the scope of the VI, there needs to be some feedback to inform the individual that their request is out of the scope of the service.

Based on her experience designing the CV website, Megan talked about the difficulties in designing for responsiveness. She believes that accommodating a large amount of information in a small device or screen is even more challenging with the localisation of CV in various languages. She also saw this as a sign that CV is growing. The varied perspectives of multiple languages, the politics of language, and locale codes or language identifiers in computing have presented interesting pain points while working on the platform. Within linguistic communities themselves, the question of locale codes for specific languages on CV is a big discussion.

4. Designing for multiple languages

While most companies and designers recognise the need for multi-language support in VI products, most VIs lack regional language compatibility. According to Akshay, the predominant languages for VI in the country are English (US), English (UK), French, and English (India). He stated that as there is insufficient data to train regional language models, the accuracy of VIs in these Indian languages will be very low. Preeti recommended that language experts must understand the technology well, and technologists must be appreciative of more diverse language models. This will enable them to collaborate better and develop higher language adaptability. The Niki team also indicated the difficulty they faced in finding the right technical expertise to build language compatibility in VI.

4.1. Using colloquial translations

Companies such as Niki and Navana Tech talked about conducting primary research on multiple-language use in languages including Hindi, Tamil, Kannada, Telugu, Oriya, Maithili, and Gujarati. The team at Niki claimed that like their technology, their design is scalable to multiple local languages. They do, however, recognise the challenges involved in adapting the app to local dialects. Before adding any new languages to Niki, the team familiarises themselves with the colloquial language used by the community in a specific region and for a specific use case. This helps them design responses according to the intent of the individuals using the app. Given the linguistic diversity of India, the biggest challenge that Niki faces is in hyper-localising conversations. The Niki team also realised from their focus group research that colloquial translations are more useful than literary ones. They aim to keep chat messages colloquial.

4.2. Support with iconography

Jai mentioned that it is challenging to incorporate speech to text as the individual may speak in a combination of English and Hindi. He shared that Navana Tech uses a combination of audio files and illustrative iconography, as this enables the individual to understand the flow of the app better. He believes that audio files allow for more realistic engagement and are a near-necessity until text to speech becomes more natural-sounding. This also helps in language accessibility.

4.3. Crowdsourcing voice data

CV has attempted to address the lack of language data by crowdsourcing it. In her interview, Megan explained that CV began with English as the primary input language but it aims to eventually have diverse language inputs. The initial prototypes of the platform were tested in Taipei. Feedback from individuals whose first language is not English, but who wanted to

contribute to the platform, made it clear that CV must be made available in more languages. The team designed a process by which individuals could contribute in their preferred language, instead of making the platform available in several arbitrary languages. The CV interface has a simple mechanism for choosing and adding languages. Further research by the team also revealed an audience for language preservation. Currently, CV is evolving to include lesser-known languages.

The CV team observed in its iterative design process that the quality of data collected needs to be more diverse in terms of gender, accent, dialect, and language. They organised an experience workshop to ideate on how to support multiple languages and improve the quality of voice data contributions.[30] Based on their learnings, they added dedicated language pages and community dashboards to the CV interface.

Our interviews also revealed that bigger consumer-based VI companies, like Amazon, Microsoft, and Google, are also considering including Indian languages. However, it is safe to assume that since the demographic of individuals using voice assistants and similar devices is likely to be English-speaking, it is easier for technology companies to continue catering to this consumer base by focusing solely on English.

5. Designing for accessibility

5.1. Accessibility for Persons with Disabilities

While voice is considered useful for people with visual impairments and certain cognitive disabilities such as dyslexia,[31] there are several other disabilities that VI design processes do not fully account for. These include cognitive disabilities, hearing impairments, physical disabilities, and non-normative speech patterns. Most of the designers we interviewed emphasised this perspective. Akshay also added that the first VI ever launched was meant to address accessibility concerns.

Preeti, who has worked on the design of a voice-based scribe for people with visual impairments, mentioned that though the overall design process was similar to that of other VI products, the design research for the scribe project was more detailed. Her team prepared the questionnaire for their research after speaking to people with visual impairments. She mentioned the need to let go of existing biases while working on the project. The team tested the product by conducting examinations with people with visual impairments and low vision; this helped them understand all the possible interactions between the individual and the scribe. Preeti believes that voice is fundamentally useful for people with low vision and those with low digital literacy. She highlighted that so far, she has not come across further work on leveraging VIs for accessibility.

CV takes certain accessibility concerns into account when designing its platform interface. They analysed the CV website using Lighthouse,[32] an open-source, automated tool that audits for performance, accessibility, and search engine optimisation (SEO) on web pages. According to their Lighthouse score, their execution of colour contrast was not up to accessibility standards. They are now ensuring that their website matches

30 Branson, M., "Prototyping with Intention," *Medium*, accessed 10 May 2021, https://medium.com/mozilla-open-innovation/prototyping-with-intention-33d15fb147c2.

31 Nielsen, J., "Voice interfaces: Assessing the Potential," *Nielsen Norman Group*, 26 January 2003, https://www.nngroup.com/articles/voice-interfaces-assessing-the-potential/.

32 "Lighthouse | tools for web developers | google developers," *Google*, accessed 17 June 2021, https://developers.google.com/web/tools/lighthouse.

these standards.

We noted that most developers and designers do not consider accessibility when conceptualising a VI product. Many VI apps use voice alongside visuals. These could be in the form of illustrations that suggest context or iconography that communicates options. Navana Tech shared the example of a voice-based banking app that uses icons to indicate options. They observed through their primary research that people needed further direction to navigate the audio-based app. The VI, therefore, uses icons as well as audio prompts for each button. The VI product created by Slang Labs is an added voice layer on an existing touch-based app. These products work well with the visual interface, and would not be as beneficial for the visually impaired as they do not operate on voice alone. Slang Labs also mentioned that by reducing the amount of screen-based interactions, they can help people who have motor disabilities like tremors.

5.2. Access and inclusivity

Preeti emphasised the need for access, along with accessibility, for the elderly, and for people with low digital literacy. Access here is the overall inclusivity for varying groups of people while accessibility here is being referred to specifically for persons with disabilities. There are also infrastructural concerns. The CV team also stressed the importance of the quality of the dataset. Large datasets are difficult to download, and they are working towards improving access. They are also working on creating a web app version of the website, which can be accessed on phones with lower connectivity, so that contributors can use it online and offline. It is evident that designers are identifying access concerns for people with low connectivity and low experience with technology; the elderly; and rural communities. However, many of these concerns are yet to be addressed in existing VIs. They still seem to be more popular with the English-speaking, mostly urban population – which is already well-versed with technology. While this focus on access is a welcome change among designers, there are clear gaps in their understanding of accessibility needs.

5.3. Learnings from grassroot initiatives

There are also key lessons to be learnt in the area of access from initiatives such as Avaaj Otalo and Gram Vaani, which have applied VI in rural India. These services have successfully used voice to increase the penetration of mobile phones even in the context of low literacy and internet access. We must note, however, that these initiatives are simpler, deterministic applications, and the probabilistic VI products currently in use face more complex training challenges.

Avaaj Otalo is a service developed in 2008 for farmers to access relevant and timely agricultural information over the phone.[33] Their team put together a set of guidelines for researchers designing VIs for developing regions. They recommend leveraging existing systems, ideating with people using the platform, and evaluating design choices empirically.[34] Putting their guidelines into action, Avaaj Otalo integrated with existing radio programmes and switched to explicit, directive-style

33 Stanford, "Voice-based social media," *Awaaz.De*, accessed 23 June 2021, https://hci.stanford.edu/research/voice4all/.

34 Patel, N., Agarwal, S., Rajput, N., Nanavati, A., Dave, P., Parikh, T., 2009. "Experiences Designing a Voice Interface for Rural India," paper presented at *Spoken Language Technology Workshop*, 2008. 21–24. 10.1109/SLT.2008.4777830, https://www.researchgate.net/publication/224382217_Experiences_designing_a_voice_interface_for_rural_India.

prompting to avoid confusion.[35]

Gram Vaani is a social tech company founded in 2008 at the Indian Institute of Technology (IIT), Delhi.[36] One of their services, Mobile Vaani, is a social media platform for social development in rural areas.[37] Mobile Vaani uses an interactive voice response (IVR) system that allows people to call a number and leave a message about their community or listen to messages left by others.[38] The platform allows communities to discuss wide-ranging issues on culture, local updates and announcements, and government schemes, and to share other information.

To design VIs that are inclusive, it is important to ensure that the training data used covers a diverse population, so that the quality of speech recognition is improved for everyone.[39] It is important to view accessibility as beneficial for everyone and not just one sub-group.[40] Taking into account the needs of individuals with visual impairment or low vision, voice interactions should be kept brief and must allow for interruptions. The application should let the individual control the speech rate. To address cognitive disabilities, a linear and time-efficient architecture is helpful. Important information should be placed at the beginning or end, and sufficient context should be provided. For individuals with hearing impairments, the VI should provide volume control and alternatives to speech-only interactions. Designers can also consider providing transcriptions of audio files or transmission to hearing devices. Physical disabilities can be addressed by enabling the VI to capture and understand broken or varying speech. Designers must include appropriate pauses in the VI's listening. For people with non-normative speech patterns, designers must provide text-to-speech alternatives.[41]

6. Designing for privacy

One of the most common privacy concerns with VIs is that the device might be always listening and collecting data. The designers we spoke with brought up this concern as well. Many devices, especially voice assistants, use 'wake words' (for example, 'Hey Alexa' or 'Okay Google') that invoke the VI. These devices are thus always on the lookout for this wake word. This is a concern if companies start recording and storing this data on the cloud. However, most companies assure users that they place the recording on the cloud and process it only when the 'wake word' is spoken. Before that, the data is stored locally on the device. This still raises questions regarding the retention of surplus data and safeguards against leaks and sharing. One of our interviewees pointed out that many privacy implications are dependent on the design of the system architecture. For example, for Google Pixel devices, all the processing happens within the device, and no data is uploaded to the cloud.

35 Patel, N. et al. "Experiences Designing a Voice Interface for Rural India."

36 "About Us," *Gramvaani*, accessed 13 July 2021, https://gramvaani.org/?p=495.

37 "Community-powered-technology," *Gramvaani*, accessed 13 July 2021, https://gramvaani.org/.

38 "How Mobile Vaani Works," *Gramvaani*, accessed 13 July 2021, https://gramvaani.org/?page_id=15.

39 Pearl, C., "Using Voice Interfaces to Make Products More Inclusive," *Harvard Business Review*, 16 May 2019, https://hbr.org/2019/05/using-voice-interfaces-to-make-products-more-inclusive.

40 Kulkarni, M., "Digital Accessibility: Challenges and Opportunities," *IIMB Management Review 31*, no. 1 (2019): 91–98, https://doi.org/https://doi.org/10.1016/j.iimb.2018.05.009. (https://www.sciencedirect.com/science/article/pii/S0970389617301131)

41 Pearl, C., "Using Voice Interfaces." .

However, Google's intention behind this was not to safeguard privacy but to optimise the processing of data as it becomes much faster compared to cloud-processing devices. In 2019, there were reports that Apple was sharing a portion of the recordings from Siri with its contractors for quality control.[42] Following this controversy, Apple updated its policy to allow people to opt into sharing audio samples of their requests to train Siri.[43]

While discussing privacy in CV, Megan spoke about the need to be transparent about how the platform is utilising the information collected. She mentioned that the dashboard helps contributors control who can see their profiles. They can hide their visibility from others on the platform. The website has been created to be as malleable as possible when it comes to contributors' interactions with it. Contributors do not need to necessarily have a profile to contribute voice data. The terms and conditions agreement states that the data is being collected for research and that personal information in the form of their voices is being collected by the website.

Preeti affirmed that while she has not seen training data that reveals an individual's identity so far people should still be aware that their voice data is being processed and utilised. This is especially essential for people who do not have a lot of experience with technology and lack trust in digital services. Preeti explained this through the example of an elderly person who finds it difficult to trust online banking and hence would be unable to use a VI to access it. She spoke about the importance of notice and consent, transparency, and opt-outs. Another interviewee also made a critical point regarding notices. While voice-only devices are always listening, they do not indicate that they are listening, in contrast to webcams, which clearly indicate – through a flash next to the lens – that they are turned on. Likewise, there needs to be a similar kind of indication in VI devices. Companies such as Amazon and Google reassure people that their devices are not violating privacy – but they also do not provide any indications of continued listening.

Another key insight that Preeti pointed out concerned designing privacy notices for voice-based products. She highlighted that communicating a privacy notice through voice could be difficult, and suggested that in such cases, there needs to be an option to view the notice as text. Designing text-based privacy notices is a challenge due to the complexity and length of these texts. Even a textual notice is difficult to access and understand.[44] Moreover, communicating a privacy notice is difficult if the device does not have a screen or if it is entirely voice-based. Taking consent verbally is another complicated design task, as the quality of consent will vary depending on the use of voice biometrics, the quality and volume of audio input, and environmental noise.[45] Preeti called for an increase in overall sensitivity about data collection and use among people. She

42 Hern, A., "Apple Contractors 'Regularly Hear Confidential Details' on Siri Recordings," *The Guardian*. 26 July 2019 , **https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings**.

43 "Improving Siri's Privacy Protections," *Apple Newsroom (India)*, accessed 29 September, 2021, **https://www.apple.com/in/newsroom/2019/08/improving-siris-privacy-protections/**.

44 Naidu, S., "Design Concerns in Creating Privacy Notices," *CIS India*, 29 May 2018, **https://cis-india.org/internet-governance/blog/design-concerns-in-creating-privacy-notices**.

45 Sigg, S., Nguyen, L. N., Zarazaga, P. P., and Backstrom, T., "Provable Consent for Voice User Interfaces," *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2020, 1–4. **https://doi.org/10.1109/percomworkshops48775.2020.9156182** (**https://ieeexplore.ieee.org/document/9156182**)

proposed an opt-out for people who do not want their data to be used for training and development purposes, while recognising that this could lead to difficulties in sourcing data for research.

Preeti remarked on larger design patterns in the voice industry: she believes that at present, the issue of privacy does not receive enough focus, and most VI companies still do not consider privacy at the conceptualisation stage. She feels the industry is still struggling to create a working system, and so privacy concerns are relegated to the end of the design process. She asserted that early conceptualisation on privacy is extremely relevant, and design is best placed to enforce and create awareness about both accessibility and ethical practices. She also pointed out the lack of guidelines for VI design. Many designers are also not familiar with data practices, privacy policies, and data protection laws. The design of privacy notices, and other elements in the interface, do not account for transparency of data collection, storage, and use. It is imperative to place VI design practice in an ethical framework and focus on privacy and transparency while designing.

Besides the policy interventions necessary to enhance privacy in VI products and services, privacy can be addressed through the interface design of VI products.[46] Designers need to provide explicit opt-ins along with enhanced notices at the time of setting up voice features. These features should not be pre-enabled. A hard 'off' switch should be available to eliminate the possibility of the device activating at inconvenient or unintended times. Creators can use prominent visual cues to practise informed consent by notifying people when a device is on and recording.

7. The future of VI design

According to Akshay, voice is just one more modality through which we interact with devices. According to him, enhanced adaptability in a VI depends on the context of the device. VIs can be more enabling for people who are not well-versed with technology. If companies begin to give importance to multi-language support, they could boost the reach of this technology. Preeti believes that VIs are going to become more and more relevant in the future. She foresees an initial phase of possible friction, but as familiarity with the technology increases, VIs will become more mainstream. She predicts that people who are not digitally literate – such as the elderly – will become the primary users of voice. VI is also likely to reduce the accessibility gap for people with disabilities.

Megan also speculated about the future of the CV website if it were to adopt voice-activated interaction as opposed to its current touch or point-and-click-based interface. The CV team feels that it is important to enable some sort of voice detection in the website, as this will allow for recordings to be more succinct and accurate. People will then be able to donate recordings of their own voices. The team could collect voice data to tune voice recognition on the platform as well. Speaking of the future of VI, Megan observed that soon there will be a homogenisation of VIs as has been the case with visual interfaces. She mentioned that this homogenisation is already underway with the use of wake words in all voice assistants. She wonders if the open-source data on CV can make this homogenisation look different. She believes that it can allow people to compete against the idea of what voice interfaces should look like. Megan also made a critical recommendation

46 Gray, S., "Always On: Privacy Implications of Microphone-enabled Devices," *FPF*, accessed 8 June, 2021, **https://fpf.org/wp-content/uploads/2016/04/FPF_Always_On_WP.pdf**.

that designers integrate ethical practices for voice at an early stage. UI/UX has already become established, but VI is still new, and the ethical foundations can be laid early in collaboration with designers.

8. Insights and further questions

Voice is projected to be a time-saving alternative to touch-based interfaces. It is often pitched as a tool for people who cannot type or read. But present applications of VI do not demonstrably bridge the gap of access and inclusivity for marginalised and vulnerable communities. As the design processes of various VI products suggest, the homogenisation of voice-based products is already underway. It is important for design to break the 'templatisation' of interfaces and allow varying applications, formats, and structures to emerge. The absence of an inclusive and contextual design practice guideline for VI is evident in the existing VI design scenario in India.

This early stage of the technology presents an opportunity to establish an ethical design framework that focuses on inclusivity, accessibility, privacy, transparency, and openness. The focus on primary research and usability is pronounced among designers, but centring on digital rights – and not just usability – is a desperate need in design practice. Our study led us to the following critical questions on design:

- How can ethics and rights become central to design practice for VIs?
- What kind of ethical guidelines should be created for designers?
- How can design enable VIs to support multiple languages?
- How can designers be familiarised with privacy and data protection practices?
- In what ways can the design process of creating VIs focus on inclusivity and accessibility?

These and other emerging questions must inform the growing landscape of work on VI in India. It is therefore imperative that the research, design, and development of these technologies are also shaped by a sustained and meaningful engagement with these thematics, and, most importantly, with the communities that would benefit the most from these advancements in technologies.

MAPPING ACTORS

To understand the landscape of voice technologies in India, we mapped 27 voice interface developers between 2019 and 2020. We used publicly available information from the interface websites and from news reports. This infographic shows developers of voice interfaces organised based on the sectors that use their service, how they identify themselves, the underlying technology they use, the languages they provide voice services in and if they disclose in their privacy policy whether they are processing voice data or not. For example Cedex Technologies usually provide their services to businesses, especially ecommerce companies, hospitality sector etc. The technology they use is NLP and they name their voice interfaces as chatbots. Though Cedex does provide multilingual support, they do not mention the languages they provide, and they do not have a privacy policy on their website.

Mapping Actors in Voice Interface Technology

Sectors where products are used	Type of service offered & technology used	Company name & typology	Language support	Disclosure
	AI	**AMAZON** SMART SPEAKER	**2** Languages	Specifies voice data processing
	NLP	**CEDEX TECHNOLOGIES** CHATBOTS	Languages not mentioned	No Privacy Policy
	NLP	**COGNIZYR** SPEECH RECOGNITION	**18** Languages	Specifies voice data processing
	AI/NLP	**DHEE AI** CONVERSATIONAL AI	Languages not mentioned	No Privacy Policy
	Voice Technology	**ESPRESSO LABS** VOICE ASSISTANTS	Languages not mentioned	Specifies voice data processing
	AI	**FLOATBOT** VOICE BOT	Languages not mentioned	Does not specify voice data processing
	AI	**GOOGLE** SMART SPEAKER	**2** Languages	Specifies voice data processing
	Speech Recognition	**GNANI.AI** AI-POWERED VIRTUAL ASSISTANT	**6** Languages	Does not specify voice data processing
	NLP	**HAPTIK** INTELLIGENT VIRTUAL ASSISTANTS	**6** Languages	Does not specify voice data processing
		JINY ASSISTIVE UI	**5** Languages	Does not specify voice data processing
		KLOVECHEF VOICE MARKETING	Languages not mentioned	Does not specify voice data processing
	AI	**KWANTICS** VOICE ASSISTANT	Languages not mentioned	No Privacy Policy
	AI	**MANTRA LABS** MULTILINGUAL, AI & VIDEO ENABLED CUSTOMER SUPPORT BOT	Languages not mentioned	Does not specify voice data processing
	Voice Technology	**NAVANA TECH** TEXT-FREE, IMAGE-BASED AND VOICE ASSISTED TECHNOLOGY	**9** Languages	No Privacy Policy
	NLP	**NIKI AI** MULTILINGUAL & VOICE BASED	**4** Languages	Does not specify voice data processing
		NUANCE TECHNOLOGY VOICE BOT	Languages not mentioned	No Privacy Policy
	Speech Recognition	**REVERIE** VOICE SUITE	**22** Languages	Does not specify voice data processing
	NLP	**SAARTHI AI** VOICE BOT	**22** Languages	Does not specify voice data processing
	NLP	**SENSEFORTH AI** CONVERSATIONAL AI	**2** Languages	Does not specify voice data processing
	Context Conversational Clustering	**SKIT** VERNACULAR INTELLIGENT VOICE ASSISTANT	Languages not mentioned	Does not specify voice data processing
	Speech Recognition	**SLANG LABS** IN-APP VOICE ASSISTANTS	**4** Languages	Specifies voice data processing
		VOKAL VOICE NOTE	**11** Languages	Specifies voice data processing
	Speech Analytics	**VOXTA** SPEECH TECHNOLOGY	Languages not mentioned	No Privacy Policy

SECTORS

FINANCE · HOSPITALS · GOVERNMENT · OTHERS

'Others' includes companies under the telecom, hospitality, transportation, and e-Commerce sectors

TYPE OF SERVICE

BUSINESS FACING · USER FACING

LANGUAGE SUPPORT

MULTILINGUAL SUPPORT · NO MULTILINGUAL SUPPORT

DISCLOSURE

SPECIFIES VOICE DATA PROCESSING · DOES NOT SPECIFY · NO PRIVACY POLICY

THE CENTRE FOR internet & society

Research **SHWETA MOHANDAS**

Design **SAUMYAA NAIDU**

Editing **PP SNEHA, SUMANDRO CHATTAPADHYAY**

Inputs **DIVYANK KATIRA, DIVYANSHA SEHGAL**

DEFINITIONS OF TECHNOLOGIES

NLP: Natural Language Processing is the branch of AI, that is works to giving computers the ability to understand human text and language.

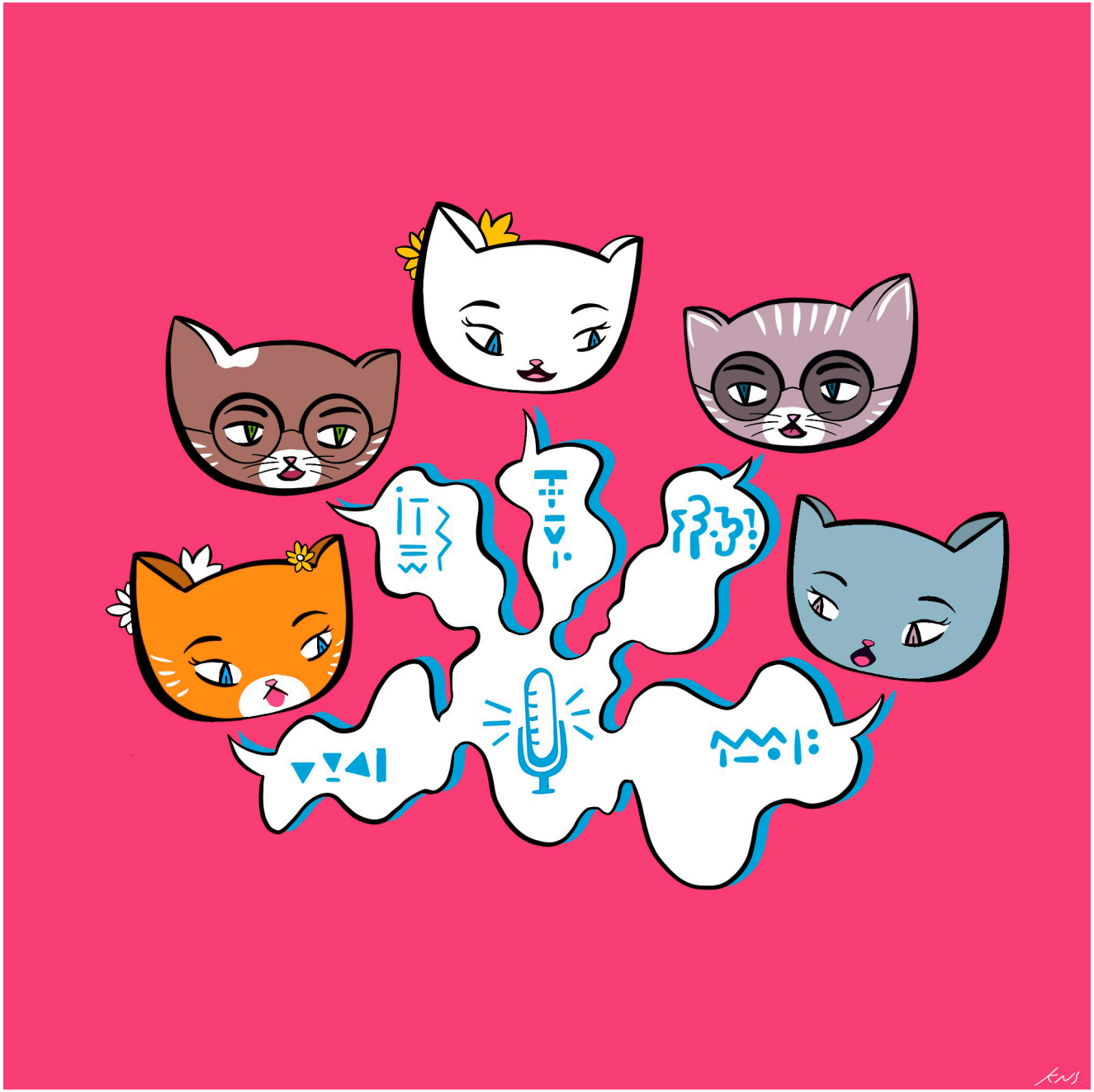
AI: Artificial intelligence is seeks to simulate human intelligence processes by machines

SPEECH RECOGNITION: Speech Recognition is the ability of a machine or program to identify words spoken to it and convert into text

SPEECH ANALYTICS: Speech analytics is the process of analysing recorded calls to gather customer information to improve communication and future interaction

CONTEXT CONVERSATIONAL CLUSTERING: Conversation Clusters attempts to bridge the verbal language barrier by using humans and machines

VOICE TECHNOLOGY: Refers to the ability of some devices to understand and respond to human speech



COMMON VOICE

Case Study

Making Voices Heard
Case Study: Common Voice

Research and Writing **SHWETA MOHANDAS, SAUMYAA NAIDU**
Review and Editing **PUTHIYA PURAYIL SNEHA, TORSHA SARKAR**
Research Inputs **SUMANDRO CHATTAPADHYAY**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. About

'... to make voice data freely and publicly available, and make sure the data represents the diversity of real people.'[1]

Common Voice (CV) is an open-source dataset of voice recordings in multiple languages that can be used to train speech-enabled applications. With over 13,905 hours of voice data across 76 different languages as of July 2021,[2] CV strives to create and maintain the largest publicly available voice dataset of its kind. CV believes that the availability of large public voice datasets will help foster innovation and create a healthy market for machine-learning-based speech technologies. In May 2020, CV began data collection for a single-word target segment (the recording of single words in multiple languages) or voice data for single-word sentences (for example yes and no), to be deployed for specific use cases or purposes. The exercise has begun with the digits zero through nine, as well as the words yes, no, hey and Firefox".[3]

2. Methodology and process

CV follows a community-driven model of creating an open-source, multilingual dataset of voice recordings that is openly accessible and usable. At the same time, it has also been working on and navigating various aspects related to privacy of voice data and accessibility for persons with disabilities, which also include complex design challenges and decisions. Some key features of this initiative include:

2.1. Community-driven contribution

"... Providing more and better data to everyone in the world who seeks to build and use voice technology."[4]

Although CV began with creating a voice dataset for English, as most of the team working on it was English-speaking, as of 2021 there are over 76 languages on the platform. CV depends on a community of volunteers and individual users who contribute voice data in order to add new languages to its website and system. One way CV promotes localisation is by localising its website to the languages it wants to add. Before adding a new language, the community has to localise 85% of the website, so that when volunteers from the local language community visit the website, they can easily navigate it, and do not need to rely on English. Then, when the language is active on the site, it is up to the community to submit 5,000 sentences that have been recorded in that language. This indicates two things to CV: a) that there is an active language community that can provide voice recordings, and b) that the barrier to including the language in CV is fairly low.

The recorded material is based on a sentence corpus that CV provides; everybody on the platform is presented with sentences that they can record and submit. These include

1 "Common Voice by Mozilla," *Common Voice*, accessed January 4, 2022, https://commonvoice.mozilla.org/en/about.

2 "Common Voice by Mozilla." *Common Voice*, accessed January 4, 2022, https://commonvoice.mozilla.org/en/datasets.

3 Branson, M., "Help Create Common Voice's First Target Segment," *Discourse*, 12 May 2020, https://discourse.mozilla.org/t/help-create-common-voices-first-target-segment/59587, 3 November 2021.

4 Roter, G., "Sharing Our Common Voices – Mozilla Releases the Largest to-date Public Domain Transcribed Voice Dataset," *The Mozilla Blog*, 9 February 2021, accessed 3 November 2021, https://blog.mozilla.org/en/mozilla/news/sharing-our-common-voices-mozilla-releases-the-largest-to-date-public-domain-transcribed-voice-dataset/

content such as parliamentary transcripts, Wikipedia articles, and sentences that members of the community have submitted. Two other community members then check to see if the audio matches the sentences. Though this is not a foolproof system, CV reports that it has a rather high accuracy rate. If people record things that are not on the card, they get voted down very quickly. This system of community curation and regulation, therefore, adds a layer of control to the accuracy and quality of content.

"Amazon and Apple, by necessity, choose languages based on what makes sense in the market and makes the most profit."[5]

Key players in the voice-as-product market serve more widely spoken languages, such as English, French, and German, because they have a large user base and hence greater demand. The issue occurs with underrepresented languages, uncommon accents, or the voices of people from underrepresented/ marginalised groups – such as those belonging to particular ethnic or gender identities. As a result, large populations remain unrepresented in datasets used to train commercial voice technologies and products. This is the gap that CV is striving to diminish.

CV's data collection differs from that of start-ups and companies like Google and Amazon; here, the sentences are self-recorded by people, and CV does not automatically detect the individual's identity, location, or other data. It does not infer the contributor's demographic based on their browsing data. Community members are also instructed not to identify people who are in the dataset.

2.2. Design process and development

Since it was envisioned as a community-driven experience, the CV team applied experience design practices when conceptualising this database.[6] Like in many design problems, the project began with the identification of a need. This need was for large quantities of publicly available voice data that could be used to train speech-to-text engines. In the design process that followed, the team ideated on creating an open-source voice dataset over the course of several design thinking exercises with Mozilla community members.[7] This resulted in paper prototypes of varying design concepts. CV then gathered in-person feedback on these prototypes to identify which design concepts to proceed on. The initial assumption of the project team was that people would need an ulterior motive to provide voice data towards this project. However, the team's insight from the research was that most people were open to the idea of voice donation. They also inferred that people wanted to learn more about the need for such voice data collection. Hence, they designed a platform whose prominent feature was collecting voice data.[8]

They developed an interactive model where people could 'teach' a robot to understand human speech by reading sentences to it.[9] This robot has become part of the CV website as a mascot

5 Interview, Common Voice, online, Bangalore, 22 October 2020

6 Branson, M., "We're Intentionally Designing Open Experiences, Here's Why," *Medium*, 10 September 2018, accessed 3 November 2021, **https://medium.com/mozilla-open-innovation/were-intentionally-designing-open-experiences-here-s-why-c6ae9730de54**, 3 November 2021.

7 Branson, "We're Intentionally Designing Open Experiences."

8 Branson, "We're Intentionally Designing Open Experiences."

9 Branson, "We're Intentionally Designing Open Experiences."

of sorts, even though the interactive teaching model is no longer operational. The alpha version of the CV platform was built "to tell the story of voice data and how it relates to the need for diversity and inclusivity in speech technology".[10] The CV team collected community feedback through tools such as Discourse[11]and Github.[12] They developed further iterations after feedback collection and discourse analysis. The Open Innovation team at Mozilla shared with us that they emphasise prototyping and reiterating. They carried out a user experience (UX) audit of the working prototype and considered community feedback from Github and Discourse. Based on this assessment, they made refinements to the platform.

Following the release of the working version, the CV team conducted another UX audit. They took into account a combination of UX heuristics, competitor evaluation (such as of platforms such as Headspace[13]), and community feedback. They looked at community feedback on Github and Discourse and spoke to the engineers who built CV. Since 2017, the focus has been on improving the platform and primarily enhancing the experience of contributing voice data. Presently, the team is looking at the bigger picture by focusing on fine-tuning the contributors' experience based on the data and research accumulated.

3. Enabling multi-language contributions

Following an iterative design process allowed CV to ask questions, derive insights, and improve its platform. The team observed that the data collected needed to be more diverse in terms of gender, accent, dialect, and language. They held an experience workshop to ideate on how to support multiple languages and enable better-quality voice data contributions.[14] They realised that the platform needed to provide people with a way to contribute in their desired language(s). They also added dedicated language pages and community dashboards. The team also made further enhancements, such as a new profile login experience and a new contribution experience, to increase the quality and quantity of voice contributions.[15]

Over the course of our interviews, we learned that CV had been designed to be a global project from the beginning. During the initial stages of development, the team ran a design sprint with a paper prototype on the streets of Taipei. It soon became clear that the platform could not be limited to English. They collected feedback from people who did not speak English as a first language, but wanted to contribute to the platform. It was evident from the feedback that CV did not need to design for specific languages, but for people to opt-in and contribute in a language of their choice. The CV interface is basic, but it features a simple mechanism to choose and add a language. Through

10 Branson, "We're Intentionally Designing Open Experiences."

11 "Civilized Discussion," *Discourse*, accessed November 1, 2021, **https://www.discourse.org/**.

12 "Where the World Builds Software," *GitHub*, accessed November 1, 2021, **https://github.com/**, 3 November 2021.

13 "Meditation and Sleep Made Simple," *Headspace*, (n.d.), accessed 3 November 2021, **https://www.headspace.com/**.

14 Branson, M., "Prototyping with Intention – Mozilla Open Innovation," *Medium*, 8 May 2020, accessed 3 November 2021, **https://medium.com/mozilla-open-innovation/prototyping-with-intention-33d15fb147c2**

15 Branson, "Prototyping with Intention."

this research, the team also discovered that there is an audience for language preservation, who wanted to add languages to CV. The team is currently looking at evolving CV for not just major languages but also for lesser-known or less visible languages.

4. Accessibility and access

The team analysed the CV website on Lighthouse,[16] an open-source, automated tool that audits web pages for performance, accessibility, and search engine optimisation (SEO). Their Lighthouse score indicated that they did not perform well in the area of colour contrast. Subsequently, they are working on ensuring that the website matches all accessibility standards. The CV team emphasised on the importance of having a high quality and accessible dataset. The files for English voice data are heavy and difficult to download, so they are working towards improving access. They are also working on creating a web app version of the website for use on devices with limited bandwidth so that contributors are able to utilise it online and offline.

5. Privacy and data collection

"We don't believe in taking information that we have not specifically been given regardless of what products are available to us." [17]

With a large number of people providing voice data, there is a need to protect privacy, especially as voices and accents are easily identifiable. As they understand the vulnerability of voice data, the CV team works closely with their trust and legal team to ensure the privacy of their contributors. They also work closely with the technical, legal, and privacy teams to ensure that the websites – and any new additions – comply with their privacy policies. Mozilla also has a data steward programme, which is run by a group of experts in the organisation who have volunteered to be consultants on data collection and best practices in data management and protection. The CV platform itself operates on two primary principles. The first is de-identification to the highest degree possible. This requires that for any language being recorded, there should be recordings by at least five people so that it becomes harder to identify them. CV also tries to remove identifiers such as sex and age in smaller datasets. The second principle is based on consent – CV does not associate voice with any client-facing data except when they consent to it. The dashboard helps contributors control who can see their profile; they can hide their visibility to others on CV. The team has created the website to be as malleable as possible when it comes to contributors' interactions with it. Contributors do not necessarily need to have a profile to contribute voice data. CV's terms and conditions agreement states that they collect data for research and that they collect personal (voice) information only when people contribute their voices.

6. Design decisions

Currently, the CV website is not voice-activated but based on 'classic' touch-/point-and-click interactions. The CV team feels that it is important to enable some sort of voice detection in the website, as this will allow for the recordings to be more succinct and accurate. The team has also been thinking about the future of the website: what would it look like when people

16 "Lighthouse | Tools for Web Developers," *Google Developers,* 2020, accessed 3 November 2021, https://developers.google.com/web/tools/lighthouse

17 Interview, Common Voice, online, Bangalore, 25 March 2020

want to donate their own voices on CV? How can CV use the data they collect to tune voice recognition on the platform itself? If they enable this, they would have to rethink the entire user experience, including navigation, actions, and initiators for contributions.

The overall objective of CV's interface design is to simplify the process by which people can contribute. It is meant to have an intuitive design. However, the experience of contributing may not be the same for everyone, and so this objective is difficult to achieve. The team observed that soon there will be a homogenisation of voice interfaces, as has been the case with websites. They note that this is already underway with wake words and voice assistants. An important question to ask here is if CV data and open-source data can make this homogenisation look different. Can they allow people to tinker and play outside of bigger entities and challenge the idea of what voice interfaces should look like?

The design team notes that it is tough to design for responsiveness. Their challenge has been to fit large quantities of information into a small device/screen, and this is exacerbated by the localisation of CV in various languages. It is difficult to design an interface where one cannot control the way the text appears across browsers. When they cannot read a language, it is difficult to troubleshoot. While this is an ongoing challenge, it is a good problem to have, as it shows that CV is growing. They affirm that taking on community feedback is the most critical and rewarding aspect of this work.

7. Challenges

A key challenge in making CV easier for contributors and the community to access is the need for internet connectivity. In addition to this, material for recording comes from sources such as parliamentary transcripts and Wikipedia, which might not reflect the actual reading and speaking styles that people use in their day-to-day lives. As these sources use more formal writing styles, the training model is also skewed towards a formal mechanism as opposed to the casual way people converse in real life. At times, women and others from underrepresented communities find it less than welcoming to engage with projects in the open-source community – including that of CV – because it mostly consists of men. This means that the dataset comprises mostly male voices, and members from diverse gender identities and communities are not adequately represented in the datasets.

8. Future of Common Voice

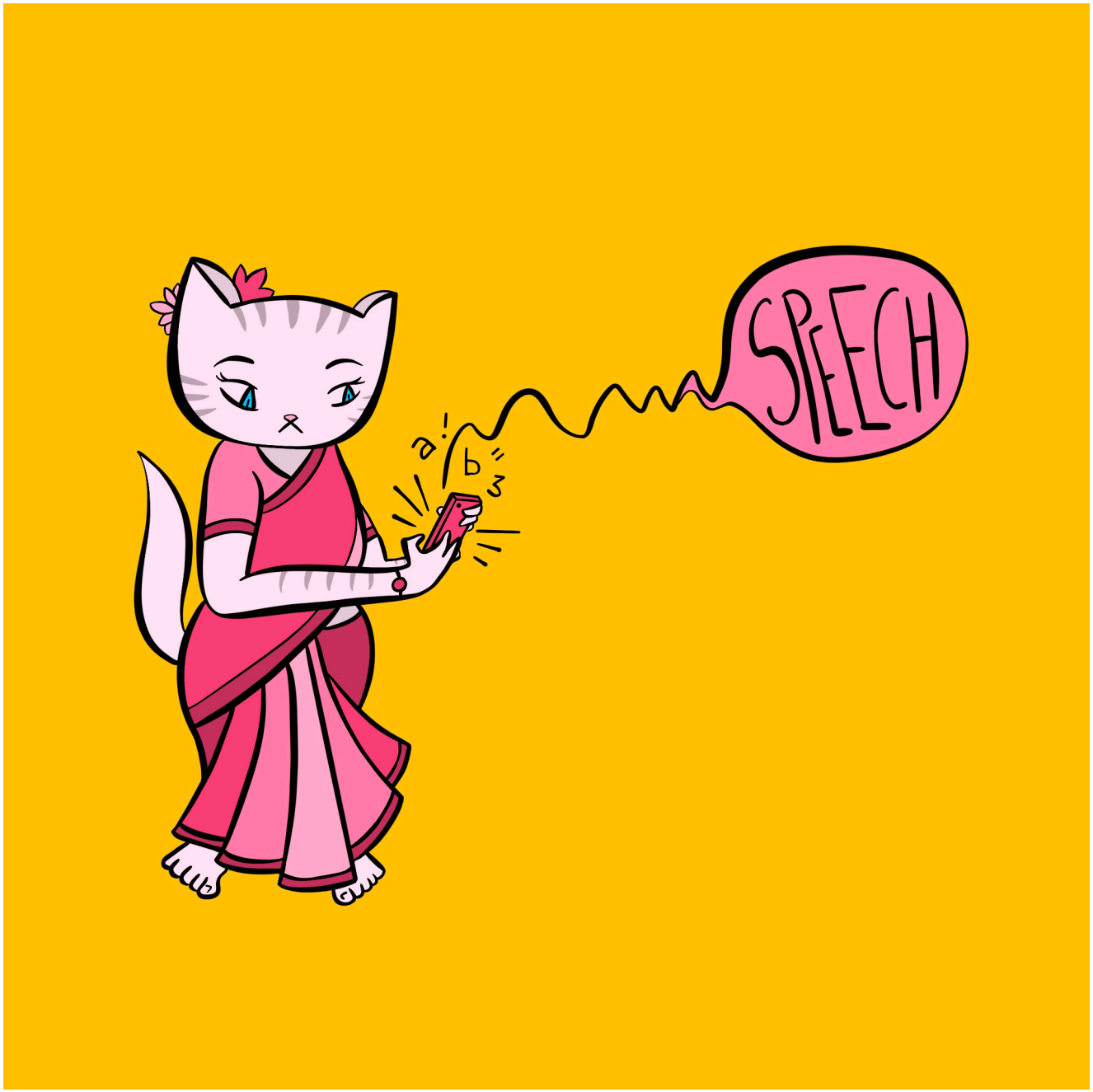
"We are only seeing an increased interest in Common Voice."[18]

CV saw a 20% growth in recorded hours during October–December 2020. Additionally, there has been a significant increase in the interest in CV, both from industries as well as communities. In recent years there has been an increase in community-driven contributions, especially from people involved in language preservation and civic duty systems. These individual and community-based initiatives help add

18 Interview, Common Voice, online, Bangalore, 22 October 2020.

more languages into the CV system, which might not have been possible with a centralised system. More recently, CV received two investments worth $1.5 million from Nvidia and $3.4 million from other investors to continue their work with native African languages.

Disclaimer: This is an independent case study conducted as a part of the Making Voices Heard Project, supported by the Mozilla Corporation. The researchers have not received any external remuneration as a part of this case study, and claim no conflict of interest.



INDIC TTS

Case Study

Making Voices Heard
Case Study: Indic TTS

Research and Writing **SHWETA MOHANDAS, SAUMYAA NAIDU**
Review and Editing **PUTHIYA PURAYIL SNEHA, TORSHA SARKAR**
Research Inputs **SUMANDRO CHATTAPADHYAY**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. About

"The amount of work in the speech domain for Indian languages is comparatively lower than that for other languages."[1]

The Indic TTS consortium was created and funded by the Department of Electronics and Information Technology, Ministry of Communications and Information Technology,[2] Government of India, to create more Indic language speech data to reduce the data divide between English and Indian languages. The Indic TTS website describes this as "a project on developing text-to-speech (TTS) synthesis systems for Indian languages, improving quality of synthesis, as well as small footprint TTS integrated with disability aids and various other applications".[3] In a recently published paper on voice technologies, researchers involved in the TTS project stated that the paucity of content in Indian languages was stark when it came to the multimedia domain and digital assistants,[4] thereby highlighting the need for speech data in Indian languages. This paucity was attributed to the lack of localisation of technologies like optical character recognition (Optical character recognition or OCR is the electronic or mechanical conversion of images of typed, handwritten, or printed text into machine-encoded text in a way that can be read by speech-to-text systems)[5],neural machine translation(neural machine translation uses computing systems that mimic the working of the human brain to predict the order of words in sentences)[6] and text-to-speech systems.

The speech data for the database was collected through the joint efforts of the 13 consortium members: IIT Madras, IIIT Hyderabad, IIT Kharagpur, IISc Bangalore, CDAC Mumbai, CDAC Thiruvananthapuram, IIT Guwahati, CDAC Kolkata, CDAC Pune, SSNCE Chennai, DA-IICT Gujarat, IIT Mandi, and PESIT Bangalore. The database and text-to-speech synthesisers were built for 13 languages, namely, Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Odia, Rajasthani, Tamil, and Telugu.

2. Methodology and process

2.1. Language and text selection

The process of creating voice datasets in Indian languages involved several steps, beginning with the selection of languages that are the focus of the project and then building speech technologies using the voice datasets. The selection of the 13 languages was based on the following criteria: optimal text selection, speaker selection, pronunciation variation, recording specification, text correction for handling out-of-the-vocabulary

1 Baby, Arun et al., "Resources for Indian Languages", In *Proceedings of CBBLR workshop, International Conference on Text, Speech and Dialogue*. Springer, 2016.

2 "Indic TTS", *Indic TTS*, **https://www.iitm.ac.in/donlab/tts/**; Department of Electronics and Information Technology (DEITY) has been renamed to Ministry of Electronics and Information Technology (MEITY) 03 November 2021.

3 "Indic TTS", *Indic TTS*. accessed 3 November 2021.

4 Baby Arun, "Resources for Indian Languages".

5 "An Introduction to Optical Character Recognition for Beginners", *Towards Data Science*, accessed 5 January 2022, **https://towardsdatascience.com/an-introduction-to-optical-character-recognition-for-beginners-14268c99d60**

6 Tan,Zhixing et al., "Neural machine translation: A review of methods, resources, and tools", *AI Open Volume 1*.(2020):5-21, **https://doi.org/10.1016/j.aiopen.2020.11.001**.

words, and data verification.[7] To ensure the quality of data, characteristics that affect speech synthesis quality such as encoding (converting one form of data to another), sampling rate (number of samples of audio recorded every second) etc. were considered. The sentences for the speech recordings were taken through web crawlers from newspaper reports, Wikipedia pages, websites, and blogs in the respective Indian language. To achieve good coverage of topics and words, sentences were also taken from different types of literature, including children's stories, science writing, tourism content, etc. Care was also taken to ensure that the texts were commonly used, free of errors, easy to read, and covered a wide range of words and syllables. Code-mixed sentences were avoided.

2.2. Speaker selection and recording

To create speech recordings for the datasets, two voice talents – a male and a female – were chosen for each language. The recordings were made in a studio room without noise or echo for clarity of the recordings. The voice talents were voice professionals who were either voice artists or newsreaders to ensure clarity in the pronunciation and diction. They were given breaks every 45 minutes to avoid fatigue. In each recording, individual sentences were isolated. A total of 40 hours of speech data was collected for a given language – 20 hours of Indian monolingual/single language data (10 hours each of male and female voice data) and 20 hours of English data recorded by first language speakers (10 hours each of male and female voice data). The recorded files were stored in .wav format to ensure that the recordings were of high sound quality.

2.3. Text-to-speech synthesis

One of the researchers in the Indic TTS project defines text-to-speech synthesis as the "process of converting an arbitrary input text to its corresponding speech output". In the context of Indian languages, the TTS system uses syllables or phonemes (units of sound that can distinguish one word from another in a particular language) as a sub-word unit (where words are split into smaller words that occur more frequently). The three major components involved in building a TTS system are text parsing, speech segmentation, and speech modelling.[8] Simply put, the objective of a TTS system is to convert text into speech output. TTS systems can be divided into two types – domain-specific and vocabulary independent. In the case of domain specific systems, the words/text to be synthesised should be limited to a particular domain, such as banking or railway broadcast, while for vocabulary independent systems, any text will work.

3. Languages

The main motivation for the project was to address the unavailability of voice data in Indian languages. The functioning of a TTS system is dependent on the training data that is fed into the system, which includes speech .wav files along with a transcript of the corresponding text. The TTS project aims to develop text-to-speech synthesisers for 13 Indian languages, which could help researchers and developers work on Indian voice applications. One of the goals of the project is to make the voice of the text-to-speech system sound as natural and understandable as possible. The first phase of the project concentrated on three languages (3 Indo-Aryan languages

7 Indic TTS", *Indic TTS*.

8 Baby, Arun, "A Unified Approach to Speech Synthesis in Indian Languages", (MS Thesis, IIT Madras, 2019), 1–93, **https://www.arunbaby.com/assets/docs/MSthesis_2019.pdf**.

and 3 Dravidian languages), the second phase added 7 more languages to the study.

4. Access and accessibility

The TTS project was started with the idea of giving people with disabilities access to regional information on the internet, such as news reports in Indian languages. Since the consortium is a publicly funded project, the datasets and research have been made public on its website. The datasets are available free of cost to researchers – they just need to log in to the website to use them. Start-ups and businesses that want to use the data can sign a Memorandum Of Understanding with Indic TTS and access the data.

5. Privacy and data collection

As stated earlier, the text data for training the systems was taken from publicly available sources such as online news portals, Wikipedia pages, websites, and blogs; hence, privacy and data protection are not significant concerns. Additionally, with regard to the speech data, the readings were done by professional voice artists who recorded sounds and words for the project based on a script provided to them by the researchers.

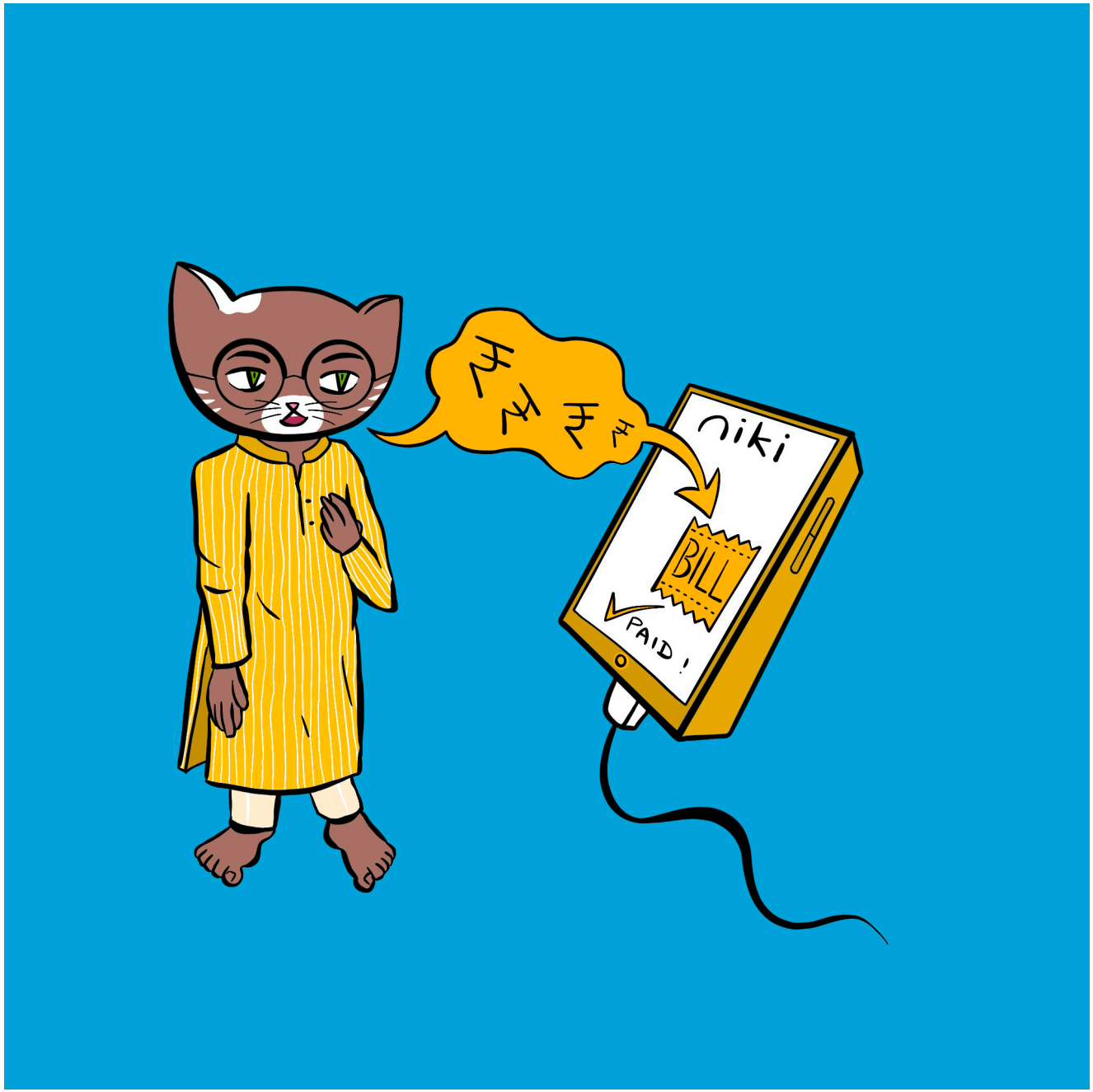
6. Challenges

One of the main challenges for the researchers was ensuring that the datasets were comprehensive and accurate while keeping the cost of creating and accessing them low. Since the project was publicly funded, the researchers needed to work with the available funding and ensure that the research was accessible and free. As stated earlier, the data and the research are open to researchers, and start-ups can request the data after signing an MOU. Another challenge was making the speech output sound more human-like and less robotic, similar to the heavily funded and data-rich interfaces of Amazon and Google. The other challenge was making the output speech systems context-specific, such as with children's books.

7. Future of Indic TTS

The project looks at continuing research and data collection with the help of government funding. Given the scale and amount of funding needed for such projects, including the requirement of infrastructure and trained human resources, the government is the primary source of funding. With the new funding from the Ministry of Electronics and Information Technology, the researchers at IITM have started a project to make English lecture videos available in Indian languages. The objective of this project is to make lectures in different domains, like humanities, healthcare, etc., freely accessible to students in their languages. This is a small-scale project, and Indic TTS hopes to expand it to more languages and subjects.

Disclaimer: This is an independent case study conducted as a part of the Making Voices Heard Project, supported by the Mozilla Corporation. The researchers have not received any external remuneration as a part of this case study, and claim no conflict of interest.



NIKI

Case Study

Making Voices Heard
Case Study: Niki

Research and Writing **SHWETA MOHANDAS, SAUMYAA NAIDU**
Review and Editing **PUTHIYA PURAYIL SNEHA, TORSHA SARKAR**
Research Inputs **SUMANDRO CHATTAPADHYAY**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. About

"Even before you do anything Niki starts to talk to you."[1]

Several start-ups in India work at the intersection of voice and multilingual support. Niki is one of the few that directly provides this support to the individual. It is described by its developers as an "artificial intelligence powered conversational commerce startup".[2] This means that a person can use the app to make an array of digital payments with either text or voice commands.[3] Niki was founded in 2015 as a voice-based assistant app in English; now, it provides information in multiple Indian languages via voice and text. The app allows people to use voice commands to complete a range of tasks, including paying utility bills, recharging prepaid mobile accounts, booking tickets for travel and accommodation, and availing local deals. The team at Niki focuses on the segment they call 'middle India', which includes customers from Tier 2, Tier 3, and Tier 4 cities; their aim is to bring the benefits of the online economy, without the barriers of language, to these new internet users. In 2020, Niki witnessed an increase of 1,000% in revenue, with a 22% increase in their user base to 550,000 users.[4] The company has received funding from Ratan Tata, Unilazer Ventures, and SAP.iO, among others.[5] As per a 2020 report, Niki plans to raise USD 50 million (more than 300 crore) by early 2021. The team intends to use these funds to expand their market share and capture 20% of the 150 million 'Bharat household' (Indian household) market by the financial year 2022.[6]

2. Methodology and process

This section enumerates the design and user research process that the team at Niki follows to understand the needs of their users and find new ways to help them.

2.1. Design process and user research

According to the team at Niki, between 2016-2019, they have spent about 30,000 hours speaking to users. Research is a critical component of their design process. They have an in-house customer insights and research team that works towards understanding users better through various methods including on-ground research. For instance, the team of researchers, designers, and product managers travelled to Tier 3 (and below) cities such as Chomu, Pushkar, Ajmer, and Udaipur in Rajasthan to meet and undertake usability studies with residents. Usability studies are conducted to observe and understand the needs of

1 Interview, Niki, online, Bangalore, 24 July 2020

2 Sangani, P., "Hindi Users Help Niki.ai Grow 300x", *The Economic Times*, 5 June 2019, https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/hindi-users-help-niki-ai-grow-300x/articleshow/69940200.cms. Accessed 05 September 2021.

3 Sangani, P., "Hindi Users Help Niki.ai Grow 300x", *The Economic Times*, 5 June 2019, https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/hindi-users-help-niki-ai-grow-300x/articleshow/69940200.cms. Accessed 05 September 2021.

4 Kashyaap, S., "[Product Roadmap] How Ratan Tata-backed Niki.ai is Helping Bharat Users Perform Transactions", *YourStory*, 8 April 2021, https://yourstory.com/2020/08/product-roadmap-ratan-tata-nikiai-bharat-conversational-ai-startup/am

5 Sangani, "Hindi Users", *The Economic Times*.

6 HT Brand Studio, "Niki Unlocks Bharat's Internet Economy Averaging 52 Transactions per Household", *Mint*, 12 January 2021, https://www.livemint.com/brand-post/niki-unlocks-bharat-s-internet-economy-averaging-52-transactions-per-household-11610451780112.html.

a group of representative users.[7] They involve observing users as they attempt to complete tasks using the product.

The following are insights based on the key learnings from the user studies and the feedback collected in Tier 3 and Tier 4 cities:

Single-page and consistently guided user interface: The entire app should have a single-page user interface (UI) across services to maintain consistency. The user journey should be guided – the app should point users to the next step. Messages need to be crisp and to the point, and actionable items and messages should not be mixed. The content flows should be designed to be linear; multiple branches should be avoided.

Acknowledgement messages: The research showed that it is important to acknowledge every action of the individual, as it gives them the confidence to use the app. All the 'call to action' and actionable elements must be consolidated in a specifically identified area.

Apprehension about change: The research revealed that users were hesitant about trying new features for fear of failing. The team observed that they were diffident and nervous about trying anything they hadn't previously encountered.

3. Multilingual and voice-based technology

The founders of Niki aim to solve two structural problems: voice and vernacular. While creating the natural language processing[8] (NLP) engine for Indian languages, the team realised that sentences in most Indian languages had some words in English, creating challenges when building an interface that could understand a sentence where more than one language was used. To deal with this, they built an in-house NLP engine, with the idea that the engine could be scaled to different contexts (such as bill payment and phone recharge) and more Indian languages. To make it easier to add more languages that the system can understand, the team now only requires entries in the Natural Language Generation[9] (NLG) files. One of the other ways in which Niki was able to increase scale in adding new languages was by delaying heavy technological investments until the proof of concept was taken to users, and by making changes and upscaling based only on the feedback and success from the user research. The NLP system designed in-house provided Niki with the capacity to launch its services in different Indian languages with ease. There are three main components in the pipeline in terms of recording a response and giving an answer – a) transcribing the audio b) extracting meaning from the text, and c) responding to the individual. In the case of Niki, the first layer is provided by Google, the second layer by the NLP engine, and the third layer by dialogue management.

3.1. Designing for hyper-localisation of conversations

Niki claims to have a scalable design that can be adapted to multiple regional languages. When expanding to a new language, the team begins by understanding colloquial usage in specific regions and for particular uses. Based on their findings, they gauge users' intentions, and design the app's responses accordingly. The key challenge was to design for the hyper-

7 "Usability Testing", *Interaction Design*, accessed 3 November 2021, https://www.interaction-design.org/literature/topics/usability-testing

8 NLP is the automatic manipulation of natural language, like speech and text, by software.

9 The software process that produces a natural language output.The NGL process converts machine code into human language output.

localisation of conversations. In India, the immense diversity of languages and dialects is difficult to capture. The team sees this as a concern to be tackled in the future.

4. Privacy & data collection

One of the main challenges faced by companies and researchers working on Indian language voice interfaces (VIs) is the lack of data in several Indian languages. Niki aims to tackle this by collecting data and strengthening their NLP engine, including as many languages as possible. One of the ways in which Niki has been striving to ensure accuracy in different languages is through the annotation of data, as they always had voice as an input for data. Based on how accurate the model is in understanding a particular language, it is made available to the user. Their NLP and machine learning operations are trained to understand what the user is saying across multiple languages. They store inputs from each interaction made by the individual using the app to improve their models for various accents and dialects. This system allows them to have data with different languages, accents, volumes, pitches, word speeds, and background noises.[10]

Niki is one of the few apps in India that provides a privacy policy in English and Hindi. The team also ensures that the financial data they collect is encrypted from the start to when the data is transferred. They also do not share personally identifiable information (PII) with third parties without the consent of the user. They practice purpose limitation internally and allow access to only the data that is required for a team to work on specific tasks.

4.1. Building for Bharat

In the initial years, Niki focused on creating voice-and speech-based interfaces that catered to English-speaking Indian people. During this time, they also provided voice-related services to various banks to use as their personalised voice bots. On listening to the interactions between people and these bots, the team realised that people were trying to speak to the English-only bot in their first languages. Only when they developed and launched the same chatbot in Hindi did they realise the immense potential that an Indian language–speaking chatbot had, especially in Tier 2 and Tier 3 cities in India.

During their extensive usability studies, they examined the reasons people buy smartphones and their device preferences. People preferred certain phones because they had big screens, their peers had them, or they wanted to watch videos on them. This is reinforced by other research that stated that people speaking Indian languages rely heavily on voice searches on non-transactional platforms like YouTube and Google.[11] However, there seemed to be hesitation to use a voice-only input platform when money was involved, this could be due to the fear that money could be transferred unintentionally. In these cases, people preferred to ask for assistance from family members or friends to complete the transaction. During the user research, the team found that the main reason for this was the fear of losing money online. Hence, the app and the voice interface were designed to ensure user trust and create an experience similar to interacting with a family member.

10 hippoBrain "E41: Full-stack Ramu Kaka - A Product to Service the Bharat Market by Sachin Jaiswal & Shishir Modi", *YouTube*, 2 April 2021, **https://www.youtube.com/watch?v=bn4h9jsmfcg&ab_channel=hippoBrainhippoBrain**.

11 Interview, Niki, online, Bangalore, 24 July 2020 - Niki

5. Accessibility and assistance

Niki equates the issue of accessibility with removal of unfamiliarity, since they believe that unfamiliarity is a barrier to using new technologies or apps. The interface is made accessible with the provision of support in multiple Indian languages in a way that makes the individual comfortable; it offers a type of hand-holding while people engage with new functionalities. The interface also includes crisp replies and acknowledgement messages. Niki believes that this could lead to an improved form of accessibility – VIs in multiple Indian languages will open the "internet economy to new internet users"[12]. This includes enabling an individual to read or interact in their own language and be comfortable using the internet for more services, including some that involve monetary transactions. Niki aims to act as a guide through the individual's journey through the app by making them comfortable, initiating conversation in their languages, and confirming each utterance to ensure an accurate record of queries.

6. Challenges

"The way Hindi is spoken in Bihar is different from the way Hindi is spoken in Rajasthan."[13]

A significant challenge for Niki and most voice interface developers in India is the huge diversity of languages that are spoken in the country. This heterogeneity of languages even among dialects creates the pressure to develop interfaces that understand languages as well as dialects. Most Indian languages are still low-resource ones without enough data, which acts as a barrier to creating VIs in these languages. Consequently the lack of less widely spoken languages available in VIs prevents them from being adopted by a number of communities.

7. Future of Niki

"Niki's vision is that nobody's left behind."[14]

Niki aims to use experience from creating local language-and-voice-based interfaces, especially for 'middle India' households in smaller cities, to expand their reach in the future. They aim to make the entire internet economy accessible to a large number of Indians who are coming online for the first time. Their current focus is to provide access to essential services such as rations, electricity, and phone recharging. Their focus has always been to improve the socio-economic lives of 'middle India' households and to further this they hope to reach more people through a voice interface that they can use in their own language.

Disclaimer: This is an independent case study conducted as a part of the Making Voices Heard Project, supported by the Mozilla Corporation. The researchers have not received any external remuneration as a part of this case study, and claim no conflict of interest.

12 Interview, Niki, online, Bangalore, 24 July 2020 - Niki

13 Interview, Niki, online, Bangalore, 24 July 2020 - Niki

14 Interview, Niki, online, Bangalore, 24 July 2020 - Niki

EVOLUTION AND TYPOLOGY OF VOICE INTERFACES

Literature Surveys

Making Voices Heard
Literature Surveys: Evolution and Typology of Voice Interfaces

Research and Writing **DEEPIKA NANDAGUDI SRINIVASA, SHWETA MOHANDAS**
Review and Editing **SAUMYAA NAIDU, PUTHIYA PURAYIL SNEHA, PRANAV MANJESH BIDARE**
Research Inputs **SUMANDRO CHATTAPADHYAY**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. Background

The availability of multiple modes of interaction such as voice and gesture makes devices accessible to a wide variety of people. Voice interfaces (VI), in particular, create a level playing field for those who are limited by single-language, text-based interfaces.

Schnelle-Walka defines VIs as "user interfaces using speech input through a speech recognizer and speech output through speech synthesis or prerecorded audio".[1] In essence, VI technologies involve two processes: one converting the language to code that a computer understands, and converting the computer language back to a language that the human understands. Considering that the predominant means of input for VIs is speech, they are also known as natural language interfaces.[2]

2. Tracing the evolution of VIs

Before Siri and Alexa, we had 'Audrey', created by Bell Laboratories' Harry Fletcher and Homer Dudley, who are considered the pioneers of VIs for their groundbreaking research on speech synthesis and human speech modelling.[3] In 1952, Audrey was used for number recognition through spoken input.[4] A decade later, IBM's 'Shoebox' could not only recognise digits from zero to nine but also comprehend 16 words.[5]

In 1992, AT&T Telefonica developed a speech-to-speech prototype, VESTS (Voice English/Spanish Translator), which relied heavily on spoken language translation.[6] VESTS, a speaker-trained system that could process over 450 words, was exhibited at the Seville World's Fair in Spain. VIs have come a long way from these early prototypes to modern voice assistants, such as Alexa, Siri, Cortana, and the Google Assistant, which are now accessible to consumers worldover.[7]

One of the main reasons for the proliferation of VIs today is that since 2012 smartphones come with a built-in VI. According to a 2018 PwC survey, consumers issued voice commands most commonly on smartphones from among a plethora of voice-enabled devices.[8] Mobile phones now operate almost like 'shrunken desktops' because of their inherent operational versatility. However, the reduced screen size is the primary structural limitation of these devices. To overcome this limitation, voice has become an important input to complete tasks without having to use the touch function or type on their

1 Schnelle-Walka, D., "I Tell You Something," *Proceedings of the 16th European Conference on Pattern Languages of Programs - EuroPLoP '11*, 2011.

2 Miller, L., "Natural Language Interfaces," *Journal of the Washington Academy of Sciences 80*, no. 3 (1990): 91–115, accessed on 3 June 2020, www.jstor.org/stable/24531256.

3 Bhowmik, A. K., *Interactive Displays: Natural Human-Interface Technologies* (John Wiley & Sons, Incorporated, 2014).

4 Carbone, C., "Audrey, Sibyl, and Alice in the Technical Information Libraries," *STWP Review 9*, no. 1 (1962): 14–15, accessed on 19 June 2020, www.jstor.org/stable/43091178

5 "IBM Shoebox," *IBM Archives*, accessed on 2 November 2021 https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html

6 IBM Archives, IBM Shoebox. Retrieved from https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html

7 Tank, N., "Voice User Interface (VUI) – A Definition," *Bot Society Blog*, 2018, https://botsociety.io/blog/2018/04/voice-user-interface/

8 "Consumer Intelligence Series: Prepare for the Voice Revolution," *PwC Survey*, 2018. https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/voice-assistants.html

phones.[9] Hence, developers have now integrated cloud-based voice technologies into devices – as in the case of Amazon Echo and Google Home as well as through open-source initiatives such as Mozilla's Deep Speech, which is an open-source speech-to-text engine.[10]

3. Features of VIs

In the early 90s, researchers identified the five basic elements[11] of voice processing technologies:

1. **Voice coding**: the process of compressing the information transmitted through the voice signal to transmit or store it economically in systems of a lower capacity.
2. **Voice synthesis**: the synthetic replication of voice signals to facilitate the transmission of information from machine to human.
3. **Speech recognition**: the extraction of information that is there in a voice signal to control the actions taken by the device in response to spoken commands.[12]
4. **Speaker recognition**: the identification of voice characteristics for speaker verification. This process ensures that the speaker is verified through their voice characteristics.
5. **Spoken language translation**: On recognising the language the person is speaking in, the translation of a message from one language to another. Through this process, two individuals who do not speak the same language can communicate.[13]

Voice output is of two distinct categories: pre-recorded speech and synthetic speech.[14] Pre-recorded speech is natural speech that is recorded and stored for future use. In contrast, synthetic speech employs natural language processing (NLP) for the automatic generation of appropriate natural-language responses or output in the form of written text.[15]

NLP involves the conversion of textual information into speech and vice-versa, which enables a device to discern and process natural language data. The system then processes this data by standardising text inputs and splitting it into words and sentences. Then, the device can ascertain the syntax of the

9 Breen, A., et al., "Voice in the User Interface," in *Interactive Displays: Natural Human-Interface Technologies*, ed. Bhowmik, A. K. (John Wiley & Sons, Incorporated, 2014): 107.

10 Lawrence, H. M. "Beyond the Graphic User Interface," In *Rhetorical Speculations: The Future of Rhetoric, Writing, and Technology*, ed. Sundvall, S., (Logan: University Press of Colorado, 2019).

11 Rabiner, L. R., "Voice Communication Between Humans and Machines –An Introduction," in *Voice Communication Between Humans and Machines*, ed. D. B. Roa and J. G. Wilpon (The National Academies Press, 1994), https://doi.org/10.17226/2308.

12 Rabiner, "Voice Communication between Humans and Machines."

13 "Voice User Interfaces," *Interaction Design Foundation*, https://www.interaction-design.org/literature/topics voice-user-interfaces.

14 Candace Kamm, "User Interface for Voice Applications", in *Voice Communication Between Humans and Machines*, eds. David B. Roe and Jay G. Wilpon (The National Academies Press, 1995), 428-429.

15 Androutsopoulos, I., *Exploring Time, Tense and Aspect in Natural Language Database Interfaces*, (John Benjamins Publishing Company, 2002).

input provided. NLP comprises two main natural-language principles:[16]

1. **Natural language understanding (NLU)**: NLU is a branch of NLP that deals with reading comprehension, synonyms, themes, and lexical semantics. It is used to construct the responses of VIs through responses of VIs algorithms.[17]
2. **Natural language generation (NLG)**: The first step of NLG involves processing relevant content from databases. This is followed by sentence planning, which involves the formation of natural-language responses through text realisation. As a consequence, the NLG process delivers a meaningful and personalised response, as opposed to a pre-scripted one.[18]

Synthetic speech employs NLP for its characteristically high 'segmental intelligibility' – or its ability to understand each segment of speech. However, pre-recorded speech outputs tend to be preferred by all for their human voice and pronunciation characteristics. These characteristics exist on the condition that the pre-recorded speech maintains the delicate balance between natural prosody[19] and the recorded elements. Since it successfully maintains the quality of natural speech, the natural prosody of pre-recorded speech output is higher than that of synthetic speech.[20]

4. Types of VIs

A plethora of developers are creating VIs that can perform various functions, thereby giving a wide array of definitions to similar interfaces. Interactive voice response (IVR), voice channels, voice bots, and voice assistants are variations of voice-based customer service solutions.[21] Although these terms are sometimes used interchangeably, some authors opine that there are nuanced differences that set them apart.[22]

4.1. Interactive voice response (IVR)

IVR systems are one of the oldest VIs in public use. These do not require a smartphone and are still used in several domains. Corkey and Parkinson (2002) define IVR as "a telephone interviewing technique in which the human speaker is replaced by a high-quality recorded interactive script to

16 "AI – Natural Language Processing", *Tutorials Point* accessed on 11 November 2021, https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm.

17 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*, 24 December, 2019, https://tinyurl.com/mrxp9emu.

18 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

19 Lauren Applebaum, et al. (2015) note that "Prosody, the intonation, rhythm, or 'music' of language, is an important aspect of all natural languages. Prosody can convey structural information that, at times, affects the meaning we take from a sentence." In "Prosody In a Communication System Developed without a Language Model," *Sign language and linguistics vol. 17*, no. 2 (2014): 181–212, doi:10.1075/sll.17.2.02app

20 Kamm, "User Interface."

21 Caile, C., "Keto or Atkins? IVR or Voice bots?" *Nuance*, 2019, accessed on 4 January 2022, https://whatsnext.nuance.com/enterprise/voice-bots-and-ivr-similarities/

22 Ghanchi, J., "Chatbots vs Virtual Assistants: Right Solution for Customer Engagement," *Medium*, 22 October 2019, accessed on 4 January 2022, https://chatbotsjournal.com/chatbots-vs-virtual-assistants-right-solution-for-customer-engagement-17fd1b06f152; Joshi, N. (2018, December 23). "Yes, Chatbots and Virtual Assistants are Different!" *Forbes*, 23 December 2018, https://www.forbes.com/sites/cognitiveworld/2018/12/23/yes-chatbots-and-virtual-assistants-are-different/#6b41450b6d7d

which the respondent provides answers by pressing the keys of a touch telephone (touch-phone)."[23] The recorded scripts used single voices,[24] combinations of male and female voices,[25] combinations of many female voices speaking in different languages,[26] or synthetic voices.[27]

4.2. Chatbots

The terms voice bots, chatbots, and automated conversational interfaces are used synonymously. They are enhanced by AI, NLP, and machine learning.[28] The term 'voice bot' is shorthand for 'voice robot'.[29] Here, voice is the primary medium of input.[30] They use automated speech recognition (ASR) technology to convert input into text. 'Chatbot' has a wider connotation, as it allows people to provide inputs in the form of text, gesture, touch, and voice. In this section, we use the term chatbot in the context of voice-enabled chatbots. The chatbot's output may be in the form of written text or voice, for which it uses text-to-speech (TTS) technology.[31] Voice chatbots can be further classified into two major categories: task-oriented (declarative) chatbots and data-driven (predictive or conversational) chatbots.[32]

a. Task-oriented chatbots

Task-oriented chatbots, also referred to as 'linguistic-based' or 'rule-based' chatbots, are devices that employ VIs that focus on a single purpose.[33] Due to this characteristic, they are considered to lack flexibility of functionality. They generate automated, conversational responses using NLP and logic. The functions of these chatbots are fairly limited, and hence they are used for specific purposes. A common example of these chatbots is interactive FAQs.

b. Data-driven chatbots

Data-driven chatbots, also known as machine-learning or AI chatbots,[34] are enhanced with AI, NLP, NLU, and machine

23 Corkrey, R., Parkinson, L., "Interactive Voice Response: Review of Studies 1989–2000," *Behavior Research Methods, Instruments, & Computers* 34 (2002): 342–353, https://doi.org/10.3758/BF03195462.

24 Piette, J. D., Weinberger, M., and McPhee, S. J., "The Effect of Automated Calls with Telephone Nurse Follow-Up on Patient-Centered Outcomes of Diabetes Care: A Randomized, Controlled Trial," *Medical Care* 38 (2000): 218–230.

25 Baer, L., Jacobs, D. G., Cukor, P., O'Laughlen, J., Coyle, J. T., and Magruder, K. M., "Automated Telephone Screening Survey for Depression," *Journal of the American Medical Association*, 273 (1995): 1943–1944.

26 Tanke, E. D., and Leirer, V. O., "Automated Telephone Reminders in Tuberculosis Care," *Medical Care* 32 (1994): 380–389.

27 Meneghini, L. F., Albisser, A. M., Goldberg, R. B., and Mintz, D. H., "An Electronic Case Manager for Diabetes Control," *Diabetes Care* 21 (1998): 591–596.

28 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

29 Middlebrook, S., and Muller, J. "Thoughts on Bots: The Emerging Law of Electronic Agents," *The Business Lawyer* 56, no. 1(2000): 341–373, accessed on 12 June 2020, www.jstor.org/stable/40687980

30 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

31 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

32 "What Is a Chatbot?", *Oracle*, accessed on 21 June 2020, https://www.oracle.com/solutions/chatbots/what-is-a-chatbot/.

33 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

34 "Chatbots: The Definitive Guide (2020)", *Artificial Solutions*.

learning to deliver personalised and meaningful responses. They are considered more interactive and contextually aware than rule-based chatbots, as their functioning is more complex and predictive.[35] This is because they learn the individual preferences and consequently create a profile of the person based on the data received.

Some refer to these types of bots as 'virtual assistants'.[36] However, other literature argues that these bots can be distinguished from virtual assistants.

4.3. Virtual assistants

According to scholars, there is no standardised definition of virtual personal assistants.[37] They list several names that other scholars have given to these systems, such as virtual assistants; vocal social agents or digital assistants; voice assistants; intelligent agents; and interactive personal assistants. Virtual assistants such as Siri use the speaker's voice and content and process it to respond in different contexts, like tasks to be performed or an action directed towards the person.[38] Virtual assistants are now increasingly used in several areas of everyday life; some common names are Siri, Google Now, Microsoft Cortana, Amazon Echo, and Google Home. These assistants interact with people in a conversational manner, thereby providing them with a wide range of functionalities.[39]

The conundrum in using the terms 'chatbot' and 'virtual assistant' interchangeably comes from the lack of universally accepted definitions. Some opine that they come under the umbrella term 'chatbots', and, in specific, 'data-driven chatbots'; the opposing view is that a virtual assistant is a completely different branch in the typology of VIs. These dissenting approaches come about because chatbots are characterised as data-obtaining interfaces.[40] In contrast, 'virtual assistant' is a distinct classification, as it is considered better than a chatbot with respect to understanding the context and the request, proficiency, nature of responses, and the rendering of a personalised experience.[41]

5. The future of VIs

VIs are slowly becoming more accessible as they are being integrated into cheaper mobile phones. The next stage is the development of smart devices for homes that can work with voice assistants, such as Google Home and Amazon Echo. On the business side, voice bots could be used for more complex customer questions. Interestingly, researchers have now also built a prototype linked with Alexa, to provide farmers with a

35 "What Is a Chatbot?", *Oracle*.

36 "What Is a Chatbot?", *Oracle*.

37 Timo Strohmann, et al., "Virtual Moderation Assistance: Creating Design Guidelines for Virtual Assistants Supporting Creative Workshops", *PACIS 2018 Proceedings*, no. 80 (2018), https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1079&context=pacis2018

38 Sirbi, K., Patankar, A. J., "Personal Assistant with Voice Recognition Intelligence", *International Journal of Engineering Research and Technology* 10, no. 1 (2017): 416–419.

39 Breen, et al, "Voice in the User Interface.".

40 Ghanchi, "Chatbots vs Virtual Assistants."

41 Joshi, "Yes, Chatbots and Virtual Assistants Are Different!"

'smart irrigation voice assistant'.[42] Similarly, a voice application named 'Avaaj Otalo' was launched by UC Berkeley School of Information, Stanford HCI Group, IBM India Research Laboratory and Development Support Center (DSC), an NGO in Gujarat, to help farmers with agriculture-related queries.[43] Lastly, another significant use of VIs, according to Joshi and Patki (2015), is in increasing the safety of the computer system. Passwords set for systems via keyboards can be duplicated. However, when it comes to securing systems via VIs, duplication becomes far more difficult.[44]

6. Conclusion

The reduction in smartphone prices and data, as well as the increase in the functions that they can perform, have enabled the integration of VIs far more complex than IVR systems. One can hope that with further data and research, there will be an increase in not just their variety, but also in their ability to communicate with people who speak different languages.

42 Ramakrishnan, V., "How Mindmeld Is Used to Conserve Agricultural Water (... and Win Hackathons in the Process)", 2019, accessed on 2 November 2021, https://www.mindmeld.com/20190828-how-mindmeld-is-used-to-conserve-agricultural-water.html

43 Patel, N., et al., "Avaaj Otalo – A Field Study of an Interactive Voice Forum for Small Farmers in Rural India." Conference on Human Factors in Computing Systems – Proceedings 2 (2010): 733–742, 10.1145/1753326.1753434.

44 Joshi, P. and Patki, R., "Voice User Interface Using Hidden Markov Model for Word Formation," *International Journal of Computer Science and Mobile Computing* 4, no. 3 (2015): 720-724, https://ijcsmc.com/docs/papers/March2015/V4I3201599a81.pdf.

VOICE INTERFACES AND ACCESSIBILITY

Literature Surveys

Making Voices Heard
Literature Surveys: Voice Interfaces and Accessibility

Research and Writing **DEEPIKA NANDAGUDI SRINIVASA, SHWETA MOHANDAS**
Review and Editing **SAUMYAA NAIDU, PUTHIYA PURAYIL SNEHA, PRANAV MANJESH BIDARE**
Research Inputs **SUMANDRO CHATTAPADHYAY**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. Background

The World Wide Web Consortium's (W3C) Web Accessibility Initiative provided a set of guidelines in 2008 and 2018 to make the internet more accessible. It also laid down the essential components of web accessibility, one of which is assistive technologies. This includes screen readers, alternative keyboards, switches, and scanning software.[1] Before the advent of consumer voice technologies, the most popular speech-based accessibility technologies were screen readers, which provide audio output for people with visual impairments, and speech dictation software, which provide a text-entry alternative to the keyboard.

The development of voice-enabled products provides the individual with the opportunity to apply speech inputs to more than just text dictation and screen reader software.[2] There has been very little research on how these applications can help people with various accessibility needs. The paucity of research could be due to the lack of funding or lack of interest from companies. Most accessibility studies focus on older adults and features such as emergency services, health monitoring, and light or temperature control. However, there has been little attention paid to how these technologies can be useful to persons with accessibility needs to perform different tasks. Despite the paucity of research, user reports show that voice-enabled devices and smart home appliances are being used by persons with disabilities (PwDs) to navigate their day-to-day activities based on speech inputs. This article explores how effective these technologies are as accessibility devices. For example, an interviewee in this study pointed out that these interfaces can be used to perform simple tasks (like relaying news or the weather) or to access entertainment (turning on YouTube or a music app), but were not effective in productivity apps (such as email dictation).

From an accessibility perspective, the adoption of voice interfaces (VIs) varies depending on the type of disability. Scholars opine that VIs were picked up as assistive technologies by persons who are visually impaired.[3] However, a significant challenge in the adoption of VIs as assistive technologies is its inability to assist deaf and hard of hearing (DHH) individuals.[4] In addition to the DHH community, older people who have debilitating conditions such as dementia,[5] and people who have speech disabilities,[6] find VIs difficult to use.

Despite these perceived limitations, Pradhan, A., et al. (2018)[7] conducted a qualitative study to ascertain the accessibility

1 Initiative, W. W. A., "Essential Components of Web Accessibility," Web Accessibility Initiative (WAI), accessed 3 November 2021, 2018, **https://www.w3.org/WAI/fundamentals/components/**

2 Initiative, "Essential Components of Web Accessibility."

3 Brewer, R., et al., "Accessible Voice Interfaces," *CSCW 18 Companion*, 2018, 441-446, accessed on 2 November 2021, **https://doi.org/10.1145/3272973.3273006**.

4 Rodolitz, J., et al. "Accessibility of Voice-Activated Agents for People Who are Deaf or Hard of Hearing," *Journal on Technology and Persons with Disabilities*, no. (2019): 144-156 .

5 Wolters, M., et al., "Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia," *Health Informatics Journal 22*, no. 4 1-13 (2015): DOI: 10.1177/1460458215593329

6 Brewer et al, "Accessible Voice Interfaces."

7 Pradhan, A., et al., "'Accessibility Came by Accident': Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities," *CHI*, 2018, **https://doi.org/10.1145/3173574.3174033**

of VIs among PwDs[8] with visual, locomotor, and cognitive impairments. The researchers examined 346 customer reviews of off-the-shelf digital assistants, such as Amazon Echo, Echo Dot, and Tap, and found that 85.6% of them were positive. The study highlighted that the people were using the Amazon Echo not just for the known uses but also for unexpected purposes such as speech therapy and support for caregivers. However it also emphasised that the people faced some difficulty in discovering new features, as well as wished for a better voice-only application.

The next section seeks to provide a holistic overview of the opportunities and challenges that individuals with disabilities face while using VIs.

2. Accessibility benefits and concerns

VIs can be beneficial to individuals who face difficulty in using text-only interfaces. Although there are multiple benefits of using VIs for performing simple to complex tasks, there is a need to look at creating devices that are universally accessible. This section will look at the benefits and concerns that come with deploying VIs as an accessibility feature for PwDs.

2.1. Individuals with vision impairment or low vision

VIs have made it easier for visually impaired people to perform simple, commonplace tasks to a certain extent. An empirical study of 16 participants with vision impairments revealed that there were different types of tasks that digital voice assistants could complete.[9] According to the sample, they used digital voice assistants to play music, check the weather, set a timer, and listen to the news. On the other hand, playing games, shopping online, calling contacts, playing the radio, and reading books were less common.[10] Emerging trends in smart devices and smart home appliances, such as Ambient Assisted Living (AAL),[11] offer new opportunities for people to access services through voice. According to Rashidi and Mihailidis, "AAL technologies provide help with daily activities, based on monitoring activities of daily living (ADL) and issuing reminders, as well as helping with mobility and automation"[12] Offshoots of AAL that utilise VIs include smart home technologies, mobile wearables or sensors, and assistive robotics, among others.[13] In an empirical study, Vacher, et al. (2013) shed light on the usability of AAL for the visually impaired, the elderly, and people

8 Pradhan, A., et al., "'Accessibility Came by Accident': Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities," *CHI*, 2018, https://doi.org/10.1145/3173574.3174033

9 Pradhan et al., "'Accessibility Came by Accident'."

10 Pradhan, et al., "'Accessibility Came by Accident'."

11 Research suggests that use of information and communication technologies (ICTs) in the daily living and working environment may enable older adults to stay active longer, remain better socially connected, and live more independently into old age. For more on AAL see: Grazia Cicirelli et al., "Ambient Assisted Living: A Review of Technologies, Methodologies and Future Perspectives for Healthy Aging of Population," *Sensors* 21, no. 10 (2021): p. 3549, https://doi.org/10.3390/s21103549.

12 Rashidi, P. and Mihailidis, A., "A Survey on Ambient-Assisted Living Tools for Older Adults," *IEEE Journal of Biomedical and Health Informatics*, 17, no. 3 (2013) 579-590.

13 Rashidi and Mihailidis, "A Survey on Ambient-Assisted Living Tools."

with no special needs.[14] They found that visually impaired participants favoured the adoption of smart home technologies, although they wished that they would render support for more complex tasks such as sending messages, emails or contacting emergency services.[15]

2.2. Individuals with locomotor disability

Smart home appliances with VIs have the potential to be of assistance to individuals with locomotor or sensory-motor disabilities. Presently, individuals can control electronic devices and the locks of their homes through voice commands.[16] Another possible advantage of VIs is that they can be potentially used to control wheelchairs.[17] In addition, VIs such as the 'listening keyboard' enable locomotor-disabled individuals to provide voice inputs, rather than traditional text inputs, to their smartphones and desktops.[18] Research proves that a listening keyboard offers better functionality than a graphical user interface; the former has a 63% better error rate and a typing rate that is 74% better.[19] If developed along the same lines as the 'listening keyboard', voice commands can also help people with limited mobility control their desktop cursors.[20]

2.3. Individuals who are deaf and hard of hearing

Although VIs are beneficial for visually impaired and motor-impaired people , the primary accessibility challenge is for DHH individuals. According to Fok, et al., "as automatic speech recognition (ASR) systems are largely trained using speech from hearing individuals, speech-controlled technologies are typically inaccessible to deaf users".[21] However, the outputs of some digital assistants for DHH individuals are also displayed as captions instead of voice. However, the subsequent problem with captions is that they become impossible for DHH individuals who are not literate to interact with them.[22]

A study revealed that a few individuals with hearing impairments who use digital assistants, faced difficulties in understanding voice outputs, although they did benefit from modifying the speech settings or pairing earphones.[23]

14 Vacher, et al., "Experimental Evaluation of Speech Recognition Technologies for Voice-based Home Automation Control in a Smart Home," Conference: Fourth Workshop on Speech and Language Processing for Assistive Technologies SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, Grenoble, France, (2013).

15 Vacher et al., "Experimental Evaluation."

16 Pradhan et al., "'Accessibility Came by Accident'."

17 Pacnik, G. et al., "Voice Operated Intelligent Wheelchair – VOIC", in Proceedings of the *IEEE International Symposium on Industrial Electronics* 3, (2005): 1221-1226, https://ieeexplore.ieee.org/document/1529099.

18 Manaris, B., et al., "A Listening Keyboard for Users with Motor Impairments – A Usability Study," *International Journal of Speech Technology* 5 (Kluwer Academic Publishers, 2002); 371–388.

19 Manaris et al.,"A Listening Keyboard."

20 Dai, L., et al., "Speech-Based Cursor Control: a Study of Grid-Based Solutions," ASSETS 2004 – The Sixth International ACM SIGACCESS Conference on Computers and Accessibility.

21 Fok, R., et al., "Towards More Robust Speech Interactions for Deaf and Hard of Hearing Users," ASSETS 2018, DOI: http://dx.doi.org/10.1145/3234695.3236343

22 Rodolitz et al., "Accessibility of Voice-Activated Agents."

23 Pradhan et al., "'Accessibility Came by Accident'."

In an attempt to explore alternative methods of using digital assistants, Rodolitz, et al. conducted an extensive study. The researchers considered the modality of using gesture control, as opposed to voice control, in a bid to use American Sign Language (ASL) instead of natural language in digital assistants.[24] Unfortunately, however, they found that with the current state of technology, it is unfeasible to use ASL to interact with digital assistants.[25]

To summarise, the literature on VIs suggests that accessibility is often incorporated as an afterthought in these technologies.[26] Incorporating the needs of individuals with disabilities into the UX design process is the need of the hour.

3. Policy schemes for accessibility

There is a shortage of technical communication research on the design of spoken language devices.[27] This lack of research translates to a lack of established standards, which is a major challenge in voice-enabled device accessibility. Significant policies promulgated in the global context to overcome this challenge include the United Nations Convention on the Rights of Persons with Disabilities (CRPD) and the World Wide Web Consortium (W3C).

India has also ratified the CRPD, whose Article 9(1) reads, "To enable persons with disabilities to live independently and participate fully in all aspects of life, States Parties shall take appropriate measures to ensure persons with disabilities access".[28] The CRPD Committee promotes the use of universal design, which encourages the development of products and services for all people without the need for specialised design.[29] Consequently, the Government of India formulated the National Policy on Universal Electronic Accessibility to ensure the adoption of universal design and accessibility standards in electronics and information and communication technologies (ICTs).[30]

One of the primary aims for formulating the CRPD and the National Policy on Universal Electronic Accessibility is ensuring the democratisation of technology. This will be achievable if interfaces are designed to be more accessible and inclusive.[31]

24 Rodolitz et al., "Accessibility of Voice-Activated Agents."

25 Shivashankara, S. and Srinath, S., "A Review on Vision Based American Sign Language Recognition, Its Techniques, and Outcomes," Proceedings of the 7th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2017.

26 Pradhan et al., "'Accessibility Came by Accident'."

27 H. M. Lawrence, "Beyond the Graphic User Interface", in *Rhetorical Speculations: The Future of Rhetoric, Writing, and Technology*, ed. S. Sundvall, (University Press of Colorado, 2019).

28 Pyaneandee, C. (2019). *International disability law: A Practical Approach to the United Nations Convention on the Rights of Persons with Disabilities*, (Taylor & Francis, 2019).

29 Article 2, UN General Assembly, "Convention on the Rights of Persons with Disabilities: resolution/adopted by the General Assembly," 24 January 2007, A/RES/61/106.

30 The Centre for Internet and Society and The Office of the Chief Commissioner for Persons with Disabilities, Department of Disability Affairs Ministry of Social Justice & Empowerment Govt. of India, National Compendium of Laws, Policies and Programmes for Persons with Disabilities, 2014.

31 Feenberg, A., "Democratizing Technology: Interests, Codes, Rights," *The Journal of Ethics* 5 no.2 (2001), 192–193.

4. The future of accessible VIs

VIs have the immense potential to be of assistance to persons who are limited by solely textual interfaces. With the increased uptake of smart appliances in homes and offices, we must consider the universal accessibility of devices so that people with various accessibility needs can use them with ease.

5. Conclusion

Voice-enabled products enable people to apply speech inputs and voice commands to access a variety of services. However, there is very little research on how these applications can help people with various accessibility needs. There is also a need to ensure that not only the device, but even the website, setup, and privacy policies are designed so everyone can access it. Additionally, developers and designers of both hardware and software should look at how to make the devices accessible to people with different types of disabilities; these could be through multiple channels of input and output, tactile markers, and audio feedback. This would go a long way in ensuring that commonly used technologies, including VIs and services are universally accessible to persons with diverse accessibility needs.

VOICE INTERFACES AND LANGUAGE

Literature Surveys

Making Voices Heard
Literature Surveys: Voice Interfaces and Language

Research and Writing **DEEPIKA NANDAGUDI SRINIVASA, SHWETA MOHANDAS**
Review and Editing **SAUMYAA NAIDU, PUTHIYA PURAYIL SNEHA, PRANAV MANJESH BIDARE**
Research Inputs **SUMANDRO CHATTAPADHYAY**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. Background

If we take voice interfaces(VIs) to be machines, then language is both the raw material and the final product – speech data is fed into these systems to train them, based on which they convert text to speech or vice versa. Hence, the key feature of VIs is the ability to convert human language into machine-readable language and vice versa. The four most significant technologies for enabling VIs, as listed by an Infosys Report in 2019, are text to speech (TTS), automatic speech recognition, natural language understanding (NLU), and natural language generation.[1] Apart from these technologies, Rudnicky enumerated the following factors needed to design a VI:[2]

- **Language design**: Refers to creating a 'habitable' language to enable the machine to "capture the range of expression"[3] of the individual, thereby creating a suitable spoken language from human–machine interaction.
- **Fluent interaction**: The process by which the individual deems the machine utilising VIs to be a competent interlocutor.
- **Recognition**: Speech recognition in VIs requires 'robustness'.[4] A robust VI is characterised by having standardised models to recognise speech efficiently.[5] Without this characteristic, the interface would be subject to systemic fluctuations in acoustic signals.[6] This would lead to modifications in input conditions which would minimally degrade the performance of the interface.[7] Building standardised models, thereby, would enable individuals to interact with VIs with high accuracy levels.[8]

2. Significant challenges for multilingual support

In an empirical study conducted by Dyches et al ,[9] 724 participants in Ohio were approached to assess the current state of the interactive voice response (IVR) system for non acute primary care. However, only 42% of the participants were able to finish the telephone screening. The rest were not able to complete the IVR process in the research for several reasons. One of the most significant reasons cited was not knowing

1 "Voice Interfaces", Infosys, 2019, accessed 3 November 2021, https://www.infosys.com/services/incubating-emerging-technologies/offerings/Documents/voice-interfaces.pdf.

2 Rudnicky, A. I., "The Design of Voice-driven Interfaces", In *Proceedings of the Workshop on Speech and Natural Language*, (Association for Computational Linguistics, USA, 1989), 120–124.

3 Rudnicky, A. I., *The Design of Voice-driven Interfaces*, 120.

4 Cole, R., et al., "The Challenge of Spoken Language Systems: Research Directions for the Nineties", *IEEE Transactions on Speech and Audio Processing*, 3, no. 1 (1995): 1–21.

5 Ayesha Pervaiz, et al., "Incorporating Noise Robustness in Speech Command Recognition by Noise Augmentation of Training Data", *Sensors* 20, no. 8 (2020): 2336–2337, https://doi.org/10.3390/s20082326.

6 Cole, R., et al., "The Challenge of Spoken Language Systems: Research Directions for the Nineties", 1–21.

7 Cole, R., et al., "The Challenge of Spoken Language Systems: Research Directions for the Nineties", 1–21.

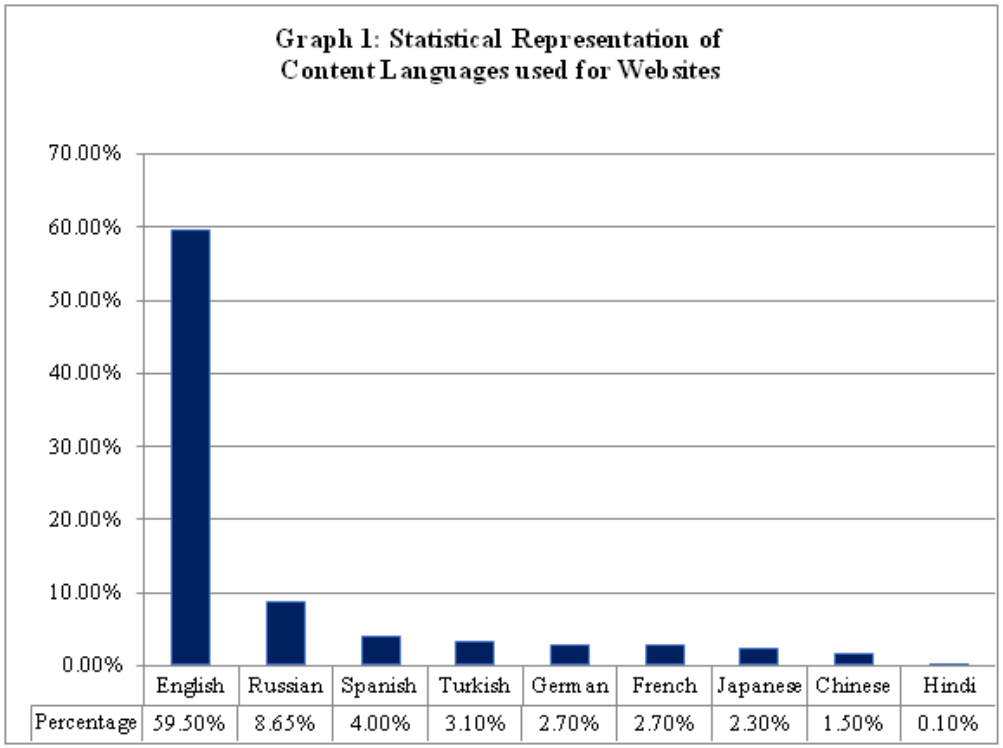
8 Rudnicky, A. I., "The Design of Voice-driven Interfaces", 120.

9 Dyches, H., Alemagno, S., Llorens, S. A., and Butts, J. M., "Automated Telephone-Administered Substance Abuse Screening for Adults in Primary Care", *Health Care Management Science*, 2, no. 4 (1999): 199–204, doi:10.1023/a:1019000231214.

English. Hence, developing a VI in all local, regional languages would be a step towards making digital spaces truly democratic.

This idea, however, has not come to fruition because of the challenge involved in developing VIs in local languages. The major challenge is further reflected in a W3Tech survey, as depicted in **Graph 1**, which reveals that English was used by 59.5% of approximately 10 million global websites as of June 2020.[10] The websites surveyed by W3Tech, however, include only websites that use technology and have "useful content". To elaborate further, default web server pages and websites owned by domain spammers were excluded from the survey. In addition, subdomains and redirected domains were not included in the survey.

The aforementioned statistics become even more significant when we consider global demographics – only 527 million people in the world, out of approximately 7.2 billion, are native English speaking people.[11] The population of native speakers[12] of three languages, namely, all Chinese dialects combined, Hindi, and Urdu is higher than the native English speaking population.[13] However, the use of these languages in website content in the two most populous countries namely China and India, are minuscule in terms of percentage. For China, it stands at 1.50%, while Hindi is behind at 0.1%. However, less than 0.1% of the 10 million (approximate value) websites surveyed accounted for using Indic languages such as Bengali, Kannada, Tamil, Telugu, Marathi, Punjabi, Gujarati, Oriya Urdu, and Assamese.[14]



	English	Russian	Spanish	Turkish	German	French	Japanese	Chinese	Hindi
Percentage	59.50%	8.65%	4.00%	3.10%	2.70%	2.70%	2.30%	1.50%	0.10%

10 "Usage Statistics of Content Languages for Websites", *W³Techs*, accessed 3 November 2021 https://w3techs.com/technologies/overview/content_language.

11 Noack, R., "The Future of Language", *Washington Post*, September 25,2015, https://www.washingtonpost.com/news/worldviews/wp/2015/09/24/the-future-of-language/.

12 The terms 'native language' and 'native speaker' are used here in the specific context of the report cited. As socio-cultural constructs, the terms have been a source of debate, particularly in postcolonial contexts and in the field of linguistics, and more recently in efforts related to language revitalisation. For more on this see: Davies, Alan. The Native Speaker: Myth and Reality. Multilingual Matters, 2003 and O'Rourke, Bernadette. "New Speakers of Minority Languages." The Routledge Handbook of Language Revitalization, 2018, 265–73. https://doi.org/10.4324/9781315561271-33.

13 Noack, R., "The Future of Language", *Washington Post*.

14 "Usage Statistics of Content Languages for Websites", *W³Techs*.

Ultimately, to address language-related challenges, building an efficient VI equipped with multilingual support is the need of the hour. This requires the expertise of computational linguists to create the domain model –i.e., build the lexicon for NLU systems and fine-tune and debug the grammar for the same.[15] Another main challenge is that it remains an expensive procedure as it requires the labour of individuals with a very niche skill set.[16] Similarly, Levinson (1994) opines that the language accessibility barriers of VIs are predominantly compounded by the lack of technical expertise to create such devices. Though the recent trend of consumer facing VIs show that there is no dearth of technical expertise, the particular nature of voice and languages still create technological challenges.

To summarise, the reluctance to develop VIs in several languages is primarily linked to the low scope for profitability and the labour-intensive requirement of computational linguists. In addition to these factors, several additional impediments have been identified for the development of interfaces in (non-dominant) local languages:

2.1. Inaccuracy

A major impediment is systemic fluctuations, which result in inaccurate speech recognition vis-a-vis natural language.[17] However, inaccuracy can be reduced by improving the interface's capability to gauge speech input with 'confidence'. A VI is deemed to be confident if it has the ability to accurately recognise even the unusual input that it receives.[18] This is predominantly in the form of words beyond the vocabulary of the interface, or different individuals interacting with the same interface, or usage of different microphones, or background noise.[19] Cole et al. opine that if a VI lacks confidence, they "produce unacceptable errors, and are unable to engage the speaker in graceful dialogues".[20] This also leads to individuals becoming frustrated with their devices due to multiple inaccurate speech interactions.[21]

2.2. Foreign accents

Inaccuracy is a challenge for the adoption of VIs, especially among non-English speaking individuals.[22] Similarly, all English speakers without an American accent tend to have significantly

15 Cole, "The Challenge of Spoken Language Systems", 1–21.

16 Cole, "The Challenge of Spoken Language", 1–21.

17 Freitas, J., et al., "Spoken Language Interface for Mobile Devices", in *Human Language Technology*. *Challenges of the Information Society*, eds. Zygmunt Vetulani, Hans Uszkoreit (Springer, Berlin, Heidelberg, 2009), 25–35.

18 "RecognizedPhrase.Confidence Property", *Microsoft*, accessed 17 November 2021, https://docs.microsoft.com/en-us/dotnet/api/system.speech.recognition.recognizedphrase.confidence?view=netframework-4.8.

19 Cole, "The Challenge of Spoken Language", 1–21.

20 Cole, "The Challenge of Spoken Language", 1–21.

21 Lawrence, H. M., "Beyond the Graphic User Interface", In *Rhetorical Speculations: The Future of Rhetoric, Writing, and Technology*, ed. S. Sundvall, (University Press of Colorado, 2019), 226-248.

22 Hernandez, Daniela, "How Voice Recognition Systems Discriminate Against People with Accents: When Will There be Speech Recognition for the Rest of Us?", *Splinter*, 21 August 2015, https://splinternews.com/how-voice-recognition-systems-discriminate-against-peop-1793850122.

less accurate interactions with VIs.[23] According to Hernandez, the error rate of VIs for American English voice interactions is 8%, with most of the words that were incorrectly identified being unique proper nouns or location names.[24] However, with Spanish and British English, the error rate was 10%.[25] The highest error rate, at 20% or above, was for the neglected 'Tier 2 languages' (languages that were not as popular with tech companies).[26] To put things in perspective, this implies that the device, on average, could not identify one out of five words spoken in a specific English accent.[27]

Like in the case of multilingual support, accent incorporation is an expensive endeavour with low chances of profitability.[28] Hence, an approach must be devised to move beyond market-driven forces to acknowledge the potential that VIs have to radically transform lives. As Lawrence rightly asserts, "as the market for speech technologies expands, the user base becomes more heterogeneous, and understanding new audiences with differing abilities, attitudes, and language backgrounds is paramount".[29]

In India, the incorporation of Indian regional languages into VIs remains a very resource-intensive task, owing to the linguistic diversity of the country.[30] Further, diverse languages have led to the emergence of different accents. Hence, this is something to be considered while using the umbrella term 'Indian accent'. Therefore, another impediment to VI adoption is the complexity involved in speech recognition for Indian accents.

Table 2 depicts the number of Indian regional languages and the 'Indian accent' supported by several voice-enabled devices. Out of the seven devices, only two supported at least one Indian language, but all seven were available in English.

23 Hernandez, "How Voice Recognition Systems" *Splinter*.

24 Hernandez, "How Voice Recognition Systems" *Splinter*.

25 Hernandez, "How Voice Recognition Systems" *Splinter*.

26 Hernandez, "How Voice Recognition Systems" *Splinter*.

27 Hernandez, "How Voice Recognition Systems" *Splinter*.

28 Lawrence, "Beyond the Graphic User Interface".

29 Lawrence, "Beyond the Graphic User Interface", 243.

30 Walkley, A. and Nagpal, J. "Why Hindi Matters in the Digital Age", *Think with Google*, 2015, from **https://www.thinkwithgoogle.com/intl/en-apac/trends-and-insights/hindi-matters-digital-age/**.

TABLE 2	
Digital assistant/voice-enabled device	**Indic language support**
Amazon Alexa	Indian English accent
Bixby	Currently does not support Indian languages
Google Assistant	English-Indian accent, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu.
Google Home	English-Indian accent
Microsoft Cortana	English-Indian accent
Microsoft Windows Narrator	English-Indian accent, Hindi, Tamil
Siri	Currently does not support Indian languages

2.3. Hybridism of English

Globalisation, transnationality, and cultural exchanges have led to the hybridism of English.[31] The most popular form in India is 'Hinglish', which is a hybrid of Hindi and English.[32] With the English language becoming the world's lingua franca, hybridism is a global phenomenon. However, despite the surge in 'Spanglish', 'Chinglish', and 'Manglish' as well as several other English hybrid forms, little to no progress has been made in developing VIs for individuals speaking in these languages.[33]

2.4. Code-switching

Code-switching is defined as speech that comprises more than one language, which is more common in multilingual communities.[34] In India, English words are often mixed into sentences in Indian languages. Researchers have found possible reasons for code-switching, such as the speaker not being able to express themselves fully in one language and switching to the other to compensate for the deficiency. Switching can also occur when an individual wishes to express solidarity with a particular social group or when the speaker tries to include people in a conversation who do not speak one of the languages.[35] In VIs and the automated processing of spoken communications, code switching presents an issue of understanding context and knowing that the added word is from a different language.

2.5. Coarticulation variability

An imperative research challenge for VIs in a linguistic context,

31 Sanchez-Stockhammer, Christina, "Hybridization in Language", In *Conceptualizing Cultural Hybridization: A Transdisciplinary Approach*, ed. Philipp Wolfgang Stockhammer, (Springer-Verlag Berlin Heidelberg, 2012), 133-157.

32 Baker, S., "Will We all be Speaking Hinglish One Day?", *British Council*, 2015, accessed 3 November 2021, https://www.britishcouncil.org/voices-magazine/will-we-all-be-speaking-hinglish-one-day

33 Lawrence, "Beyond the Graphic User Interface".

34 Skiba, R., "Code switching as a Countenance of Language Interference", *The Internet TESL Journal*, 3, no. 10 (1997): 1–6.

35 Crystal, D. *The Cambridge Encyclopedia of Language*, (Cambridge University Press, 1987), 372-375.

as observed by Cole et al. (1995), is "coarticulation variability."[36] The term refers to the inherent linguistic subjectivity of a sound segment due to factors such as accent, idiolect, and sociolect.[37] For instance, linguistic subjectivity can be observed with French, as the same language varies tremendously when spoken in France and Canada.[38]

In the Mozilla Common Voice project, the collection of voice data segments for machine learning is a two-pronged process involving contributors recording voice clips and the verification of the accuracy of the same recording.[39] If two individuals vote that the voice recording provided is accurate, it will enter the Common Voice dataset; however, if two individuals do not approve of the recording, it will enter what Common Voice terms as the 'Clip Graveyard'.[40] However, this process can be biased due to coarticulation variability – a voice recording might get sent to the Clip Graveyard if the articulation of words, despite being accurate, does not match the pronunciation of the individual verifying the recording. However, Common Voice has explicitly acknowledged this limitation vis-a-vis their voice corpus.[41]

3. Voice initiatives to bridge the digital divide

Paul (2017) opines that a possible method to resolve the linguistic limitations of VIs, is to train the device employing a VI to associate particular sounds with words.[42] Training a machine to recognise sounds requires an extensive database of voice recordings on a wide variety of topics. The flexibility and accuracy of the VI are dependent on the number of voices and accents it is exposed to.[43] Presently, several eminent organisations, universities, and government bodies have undertaken the challenging task of creating such extensive voice databases:

3.1. Global initiatives

In an attempt to create a database to foster the growth of inclusive technologies, the Mozilla Foundation launched the Common Voice project in 2017.[44] To facilitate machine learning vis-a-vis VIs, developers require a large amount of voice data, which is usually expensive and resource-intensive to collect. Hence, Common Voice encourages people to donate their voice

36 Cole, "The Challenge of Spoken Language", 1–21.

37 Martin, R., "Common Voice Languages and Accent Strategy v5", *Mozilla*, 2020, accessed 3 November 2021, **https://discourse.mozilla.org/t/common-voice-languages-and-accent-strategy-v5/56555**

38 McEvoy, J., "A Few Differences Between French Spoken in Québec and France", *British Council*, 2017, accessed 3 November 2021, **https://www.britishcouncil.org/voices-magazine/few-differences-between-french-spoken-quebec-and-france**

39 "Why Common Voice?", *Common Voice*, **https://commonvoice.mozilla.org/en/about**

40 "Why Common Voice?", *Common Voice*.

41 Martin, R., "Common Voice Languages and Accent Strategy v5", *Mozilla*.

42 Paul, S. "Voice Is the Next Big Platform, Unless You Have an Accent", *Wired*, 2017, **https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/**

43 Paul, S., "Voice Is the Next Big Platform, Unless You Have an Accent", *Wired*.

44 "Why Common Voice?", *Common Voice*.

recording samples as well as verify other voice clips, thereby creating an accurate, open-source, and truly diverse database of voices.[45] The project also recently initiated work on collecting single word segments, which aims to enable the machine to identify numbers (zero to nine) and the words 'yes', 'no', 'hey', and 'Firefox'.[46] As of July 2021, the Common Voice project had collected 13,905 hours of recordings in 76 different languages.[47]

The Linguistic Data Consortium (LDC) was conceptualised in 1992 to enhance technologies to support language-based academia.[48] LDC served as the leading language repository for educational institutions, corporations, and research institutes.[49] The repository was formed as a result of the LDC's collaborations with researchers, who are instrumental in evaluating the voice data collection. LDC also has agreements with 40 organisations to create a general corpus. One of them is Microsoft Research India, which deals exclusively with Indian language tagsets.[50] A 'tag' refers to the "labels used to indicate the part of speech", which also include the grammatical aspects of the language.[51] A 'tagset' is a collection of tags made by organisations such as Microsoft that deal with corpus creation.

VoxForge is an open speech dataset that was set up to collect transcribed speech with Free and Open Source Speech Recognition Engines (on Linux, Windows, and Mac).[52] The submitted audio files have been made available under a General Public License (GPL) license and then compiled into acoustic models for use with Open Source speech recognition engines such as CMU Sphinx, ISIP, Julius (Github), and HTK.

M-AILABS Speech Dataset is the first large dataset that is available free-of-charge and usable as training data for speech recognition and speech synthesis.[53] Most of the data is derived from LibriVox (which provides free public domain audiobooks) and Project Gutenberg (which provides free e-books). The training data consists of nearly a thousand hours of audio and text files in prepared formats.

3.2. Initiatives for Indian languages

The National Platform for Language Technology (NPLT) is a platform for colleges, researchers, and companies to provide access to Indian language data, tools, and related web services.[54] The NPLT acts as a marketplace of linguistic resources, tools and

45 "Why Common Voice?", *Common Voice*.

46 Branson, M., "Help Create Common Voice's First Target Segment", *Mozilla*, 2020, accessed 3 November 2021, **https://discourse.mozilla.org/t/help-create-common-voices-first-target-segment/59587**

47 Branson, M., "More Data, More Languages, and Introducing our First Target Segment!", *Mozilla*, 2020, accessed 3 November 2021, **https://discourse.mozilla.org/t/common-voice-dataset-release-mid-year-2020/62938**

48 "Mission", *Linguistic Data Consortium*, accessed 3 November 2021, **https://www.ldc.upenn.edu/about/mission**

49 "About LDC", *Linguistic Data Consortium*, accessed 3 November 2021, **https://www.ldc.upenn.edu/about**

50 "Other Collaborations", *Linguistic Data Consortium*, accessed 3 November 2021, **https://www.ldc.upenn.edu/collaborations/other**

51 "Tagset for Indian Languages", *Sketch Engine*, accessed 3 November 2021, **https://www.sketchengine.eu/tagset-indian-languages/**

52 "VoxForge", *VoxForge*, **http://www.voxforge.org/**.

53 "The M-AILABS Speech Dataset", *Caito*, accessed 3 November 2021, **https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/**.

54 "About Us", *National Platform for Language Technology*. **https://nplt.in/demo/about-nplt**, ¶3.

services developed by the government, start-ups, industries, and other stakeholders. The platform makes these resources available to interested entities, be it researchers, academicians, start-ups, or MNCs, for research and commercial purposes. It acts as a marketplace for Indian language data in both speech and text, with the aim of lending power to machine learning algorithms and improving the accuracy of models. NPLT also aims to provide a central point of "discoverability of Indian Language Data, technologies and services etc" to satisfy the data needs of both industry and academia.

Indic TTS is a joint initiative by the Government of India and 13 eminent Indian institutions. However, unlike Common Voice, which is a database for several languages across the world, Indic TTS focuses on 13 Indian languages.[55] The special corpus consists of over 10,000 sentences and words spoken by both male and female speakers. The Indic TTS project also successfully launched an Android application for the TTS synthesis of 13 Indian languages.[56] By utilising the unified parser, this application could recognise text input in 13 different Indian languages and render spoken output.

4. Future of multilingual VIs

Lawrence hints that the Hindi language will be the next most represented language in speech technology.[57] This is attributed to the fact that India is considered an emerging market economy.[58] Similarly, large stakeholders in the digital economy, such as Google are working on the prediction that a large percentage of next-generation Indian internet users will be Hindi speakers as opposed to English speakers. Hence, Google is now taking measures to enhance its software user interfaces and products to cater to Hindi-speaking consumers.[59] This step is incentivised by profits, but the silver lining is that it significantly addresses the 'digital speech divide' dilemma.[60]

5. Conclusion

Voice-based technologies have the potential to make the internet more accessible compared to purely text-based interfaces. What people can do with the internet can be significantly increased and improved if they can communicate in their own language. However, the need for data and the ever-changing nature of languages and their contexts can be a challenge for interfaces in multiple languages. One can hope that the push towards more voice-based interfaces and the need for language data will bring in interest and funding towards the creation of language data corpora in more languages.

55 "Voices", *Indic TTS*, https://www.iitm.ac.in/donlab/tts/voices.php

56 "Android Applications", *Indic TTS*, https://www.iitm.ac.in/donlab/tts/androidapp.php

57 Lawrence, "Beyond the Graphic User Interface".

58 Hernandez, "How Voice Recognition Systems" *Splinter*.

59 Walkley, A. and Nagpal, J., "Why Hindi Matters in the Digital Age", *Think with Google*.

60 Walkley, A. and Nagpal, J., "Why Hindi Matters in the Digital Age", *Think with Google*.

VOICE INTERFACES AND PRIVACY

Literature Surveys

Making Voices Heard
Literature Surveys: Voice Interfaces and Privacy

Research and Writing **DIVYA PINHEIRO, SHWETA MOHANDAS**
Review and Editing **SAUMYAA NAIDU, PUTHIYA PURAYIL SNEHA, PRANAV MANJESH BIDARE**
Research Inputs **SUMANDRO CHATTAPADHYAY**

CENTRE FOR INTERNET AND SOCIETY
Supported by Mozilla Corporation

Contents

1. Background

Efforts to develop technologies with voice recognition have been ongoing since the 1960s. Though significant advances in this field were seen in the 1990s with the advent of personal computers (PCs), the biggest breakthrough was the introduction of Siri (a voice-based virtual assistant) on the Apple iPhone in 2011. The use of voice-controlled technologies is not limited to mobile phones and smart speakers; now, they are also integrated with other smart devices such as PCs (as in the case of Microsoft's Cortana), TVs, and cars.[1]

However, the increased use of voice interfaces (VIs) has led to the emergence of a host of concerns, specifically surrounding user privacy. According to a 2020 study, about 33% of adults surveyed reported that privacy concerns were a top reason for not purchasing devices with built-in VI systems; this figure saw a significant increase from 16% in 2018 and 23% in 2019.[2] This article aims to analyse the privacy concerns surrounding VIs, both now and in the future.

2. The spectrum of VI devices

The wide prevalence of microphone-enabled devices today has ushered in an "era of Ubiquitous Listening".[3] VI devices can be categorised into three kinds –

- Manually activated devices – The person presses a button that causes the device to turn on and begin recording.
- Speech-activated devices – These devices remain in an inert state of passive processing. The device re-records local information without transmitting or storing any information and only begins actively recording when it detects its trigger word or 'wake word',[4] such as 'Hey Siri' or 'Ok Google'.
- Always on devices – These devices are designed to record and transmit data all the time until turned off.[5]

Privacy concerns arise, particularly, in the latter two categories, where devices can access, record, and store the data of the individual.[6] Most people do not understand when a VI is listening and where their data is being stored.[7] The data thus

1 Youval Nachum, "Privacy Issues with Voice Interfaces", *EEWeb*, 1 July 2019, https://www.eeweb.com/privacy-issues-with-voice-interfaces/.

2 Bret Kinsella, "Privacy Concerns Rise Significantly as 1-in-3 Consumers Cite It as a Reason to Avoid Smart Speakers", *Voicebot.ai*, 11 May 2020, https://voicebot.ai/2020/05/11/privacy-concerns-rise-significantly-as-1-in-3-consumers-cite-it-as-reason-to-avoid-smart-speakers/.

3 David Talbot, "The Era of Ubiquitous Listening Dawns", *MIT Technology Review*, 8 August 2013. http://www.technologyreview.com/news/517801/the-era-of-ubiquitous-listening-dawns/.

4 Stacey Grey, *Always on: Privacy Implications of Microphone-Enabled Devices*, Future of Privacy Forum, 16 April 2016, https://fpf.org/wp-content/uploads/2016/04/FPF_Always_On_WP.pdf.

5 Grey, Always on: Privacy Implications of Microphone-Enabled Devices.

6 Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, David Wagner, "Privacy Attitudes of Smart Speaker Users", *Proceedings on Privacy Enhancing Technologies*, 2019, no. 4 (2019), 250–271.

7 Nathan Malkin, Julia Bernd, Maritza Johnson, and Serge Egelman. "'What Can't Data Be Used For?' Privacy Expectations about Smart TVs in the US", *Proceedings of the Third European Workshop on Usable Security*, 23 April 2018, https://www.ndss-symposium.org/wp-content/uploads/2018/06/eurousec2018_16_Malkin_paper.pdf.

collected is often exploited for targeted advertising.[8] Patent filings at the United States Patent and Trademark Office (USPTO) indicate an increase in the development of always-on devices that listen to things beyond the device's wake word to perform increasingly sophisticated analysis.[9]

3. Primary concerns related to privacy

The same features of speech recognition that make such devices appealing are also those that give rise to privacy concerns as VIs become increasingly integrated with our daily lives.[10] Though such services typically require user permission to work, it is usually granted if people are interested in its use.[11] A survey of Android users found that only 17% of respondents paid attention to permissions during app installations and only 3% were able to answer questions on these permissions.[12] Unlike phones or devices that are used by specific individuals, VIs such as Google Home and Amazon Echo can collect data from people who have not consented to their conversation being recorded. This could include visitors, workers, and even children.

3.1. Listening in to private conversations

One of the main issues concerning voice-based virtual assistants is that the device can be activated through the accidental use of wake words. This constant listening has raised concerns regarding devices eavesdropping on private conversations as well as the processing and sharing of data with third parties including law enforcement agencies.[13]

The Supreme Court of India recognised the right to privacy as implicit in the right to life and liberty under Article 21. This includes the right to be left alone. A citizen has the right to safeguard their own privacy as well as that of their family, educational details, etc. Such information can only be published with the person's consent.[14] One of the major privacy concerns associated with the use of constantly listening VIs is that there is a high chance of third parties listening in on private conversations through the device.

These devices mostly record information on hearing the wake word. However, people may unintentionally cause the device to

8 John M. Simpson, "Home Assistant Adopter Beware: Google, Amazon Digital Assistant Patents Reveal Plans for Mass Snooping", *Consumer Watchdog*, 2017.

9 Nathan Malkin, Serge Egelman, and David Wagner, "Privacy Controls for Always-listening Devices", *Proceedings of the New Security Paradigms Workshop*, 2019, 78–91.

10 Stacey Grey, Always on: Privacy Implications of Microphone-Enabled Devices, Future of Privacy Forum, 16 April 2016, https://fpf.org/wp-content/uploads/2016/04/FPF_Always_On_WP.pdf.

11 Dan Arp, Erwin Quiring, Christian Wressnegger, and K. Rieck, "Privacy Threats through Ultrasonic Side Channels on Mobile Devices", 2017 *IEEE European Symposium on Security and Privacy*, 2017, pp. 35–47.

12 Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner, "Android Permissions: User Attention, Comprehension, and Behavior", in *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS 2012)* (ACM Press, 2012).

13 Sidney Fussell, "Police Want Your Smart Speaker—Here's Why", *Wired*, 23 August 2020, https://www.wired.com/story/star-witness-your-smart-speaker/.

14 *R Rajagopal v. State of T.N.* (1994) 6 SCC 632, pp. 649–51.

begin recording if a word similar to the wake word is spoken. A study found that more than 1,000 terms can activate VI devices, highlighting the scope of the potential risk to privacy.[15] The phrases were not just limited to those that sounded very similar to the wake words, but also remote words such as 'unacceptable' (to which Alexa was activated) and 'tobacco' (to which Echo was activated). This finding is further reiterated by the results of a study that found that VI devices were accidentally activated by 64% of people using it, in a month.[16]

Recently, it was revealed that the big five tech companies – Amazon, Apple, Facebook, Alphabet/ Google, and Microsoft – have been using human contractors to analyse a small percentage of VI recordings. These recordings, although anonymous, can potentially contain personal information, resulting in an infringement of user rights.[17] A report also found that the information passed on included sensitive personal information such as the latitude and longitude coordinates associated with the voice data, which could indicate a person's home address.[18]

Third-party access to the personal information of individuals not only raises questions regarding privacy but also paves the way for other uses of this data such as for profiling and surveillance.

3.2. Access and use of VI data by law enforcement

Digital data has become increasingly useful to law enforcement and security agencies, with the police relying on wearables and smart devices to verify the claims of people made during an investigation.[19] The first instance of the use of VI data as evidence was in a 2015 murder case in the United States, in which a man was found dead in a hot tub. Investigators issued a warrant to Amazon, requiring the company to turn over information and audio recordings captured by the suspect's Echo speaker.[20] VI devices have since been used both to exonerate[21] as well as to hold suspects guilty of crimes.[22]

15 Eric Hal Schwartz, "More than 1,000 Phrases Will Accidentally Awaken Alexa, Siri, and Google Assistant: Study", *Voicebot.ai*, 6 July 2020, **https://voicebot.ai/2020/07/06/more-than-1000-phrases-will-accidentally-awaken-alexa-siri-and-google-assistant-study/**.

16 Eric Hal Schwartz, "Voice Assistants Accidentally Awakened by 64% of Users a Month: Survey", *Voicebot.ai*, 9 January 2020, **https://voicebot.ai/2020/01/09/voice-assistants-accidentally-awakened-by-64-of-users-a-month-survey/**.

17 Dorian Lynskey, "'Alexa, Are You Invading My Privacy?' – The Dark Side of Our Voice Assistants", *The Guardian*, 9 October 2019, **https://www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants**

18 Sarah Perez, "41% of Voice Assistant Users Have Concerns about Trust and Privacy, Report Finds", *TechCrunch*, 25 April 2019, **https://techcrunch.com/2019/04/24/41-of-voice-assistant-users-have-concerns-about-trust-and-privacy-report-finds/**.

19 Fussell, "Police Want Your Smart Speaker", *Wired*.

20 "Servant or Spy? Law Enforcement, Privacy Advocates Grapple with Brave New World of AI Assistants", *CNBC*, accessed 24 November 2021, **https://www.cnbc.com/2017/01/06/servant-or-spy-law-enforcement-privacy-advocates-grapple-with-brave-new-world-of-ai-assistants.html**.

21 Kayla Epstein, "Police Think Amazon's Alexa May Have Information on a Fatal Stabbing Case", *Washington Post*, 3 November 2019, **https://www.washingtonpost.com/technology/2019/11/02/police-think-amazons-alexa-may-have-information-fatal-stabbing-case/**

22 Juang,"Servant or Spy?",*CNBC*

This risks creating a culture of state surveillance of the daily activities of citizens with potentially worrying consequences.[23] As more of such data is collected, we must ensure that it receives robust protection.[24]

3.3. Data used for advertisement strategies

VI manufacturers use the data collected from people using the devices to enhance their advertisement strategies. Patents filed in the United States reveal how these devices can be used for massive information collection and intrusive digital advertising.[25] Such data is collected on the pretext of providing customers with advertisements customised to their interests.[26] VIs greatly benefit advertisers who rely on complex data sets to make essential advertising decisions. The massive amount of data gathered from app and platform VI interactions allow for efficient processing, analysis, and access of data.[27] Although the practice is currently uncommon, and manufacturers currently have policies that specifically restrict advertisements on VI devices, there is potential for their use as a mode of advertisement that informs users of content that caters to their interests.[28]

3.4. The privacy of children on VI devices

Two-thirds of India's internet users are in the 12–29 years age group, with those in the 12–19 age group accounting for about 21.5% of the total internet usage in metro cities.[29] Children today utilise the internet to access information, education, and other opportunities.[30] The risk to privacy is one of the primary concerns pertaining to children's use of the internet. Children on the internet are less likely to have a comprehensive understanding of the consequences of privacy infringement, making them a vulnerable group that needs added protection.[31]

Chapter IV of the Personal Data Protection Bill, 2019 (PDP), lays down special conditions for the processing of a child's data. Such processing must be done with the intention of ensuring

23 Garfield Benjamin, "Amazon Echo's Privacy Issues Go Way Beyond Voice Recordings", *The Conversation*, 21 January 2020, https://theconversation.com/amazon-echos-privacy-issues-go-way-beyond-voice-recordings-130016.

24 Joseph Jerome, "Alexa, Is Law Enforcement Listening?" *Center for Democracy and Technology*, 4 January 2017, https://cdt.org/insights/alexa-is-law-enforcement-listening/.

25 John M. Simpson, "Home Assistant Adopter Beware: Google, Amazon Digital Assistant Patents Reveal Plans for Mass Snooping", *Consumer Watchdog*, 13 December 2017, https://www.consumerwatchdog.org/privacy-technology/home-assistant-adopter-beware-google-amazon-digital-assistant-patents-reveal.

26 Simpson, " Home Assistant Adopter Beware", *Consumer Watchdog*.

27 Jason Hall, "How Artificial Intelligence is Transforming Digital Marketing", *Forbes*, accessed 24 November 2021, https://www.forbes.com/sites/forbesagencycouncil/2019/08/21/how-artificial-intelligence-is-transforming-digital-marketing/?sh=39700bde21e1.

28 Jesus Martín, "Advertising in Voice Interfaces", *UX Collective*, 14 July 2020, https://uxdesign.cc/advertising-in-voice-interfaces-4b1ca14fa28b

29 Neilsen, "Digital in India 2019 – Round 2 Report", *IAMAI*, accessed 24 November 2021, https://reverieinc.com/wp-content/uploads/2020/09/IAMAI-Digital-in-India-2019-Round-2-Report.pdf.

30 UNICEF, "The State of the World's Children 2017. Children in a Digital World", 2017, https://www.unicef.org/publications/files/SOWC_2017_ENG_WEB.pdf

31 Sonia Livingstone, "Children: A Special Case for Privacy?" *International Institute of Communications*, 19 December 2019, http://www.iicom.org/intermedia/intermedia-july-2018/children-a-special-case-for-privacy.

the best interests of the child after taking appropriate steps to verify their age and on receiving the consent of a parent or guardian.[32] The European Union's General Data Protection Regulation (GDPR)[33] and the Children's Online Privacy Protection Rule (COPPA)[34] in the United States also provide similar protections. These provisions have, however, not laid down explicit consequences for non-compliance with these rules; the Federal Trade Commission in the US has been slow to impose hefty fines for such acts and the still-young GDPR has not dealt extensively with such issues.[35]

4. Voice biometrics and the future steps for voice technologies

One of the more recent advancements in this area is the use of voice biometrics to authenticate the person using the device. Voice biometrics require that the system first process a voice sample to extract speaker-specific characteristics to build a statistical model, referred to as a voiceprint or a voice signature. Following this, any new input is compared with the existing voice signature for verification.[36] Data collected through VIs may also fall within the purview of biometric data. The Supreme Court of India, in the landmark case of *Justice Puttaswamy v. Union of India*, characterised biometric data as that which is intrinsically linked to humane characteristics.[37] The Personal Data Protection (PDP) Bill classifies biometric data as sensitive personal data that requires explicit consent for processing.[38]

Voice biometrics seem to be the proposed way forward for VIs. Google has confirmed that it is working on a new Google Assistant feature that can be used to authorise financial transactions through voice biometrics.[39] Unlike identifiers such as phone numbers, address or email ids, biometrics cannot be discarded or replaced. This raises significant privacy issues relating to how such data are collected, processed, and stored. The data may be used for purposes other than that for which they were initially collected (a phenomenon also known as function creep).[40]

Recent advancements in technology pose threats to the privacy of individuals who make use of these services. This issue

32 Section 16, Personal Data Protection Bill, 2019.

33 General Data Protection Regulation (GDPR), https://gdpr-info.eu/.

34 Federal Trade Commission, "Children's Online Privacy Protection Rule", https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.

35 Martyn Farrows, "Let's Talk Voice Tech, Data Privacy, and Kids", *VoiceBot.AI*, 28 March 2020, https://voicebot.ai/2020/03/28/lets-talk-voice-tech-data-privacy-and-kids/.

36 Abhijit Ahaskar, "Voice Biometrics Are Cleverer Now, But Still Need More Work", *LiveMint*, 6 February 2020, https://www.livemint.com/technology/tech-news/voice-biometrics-are-cleverer-now-but-still-need-more-work-11581011267941.html.

37 *K.S. Puttaswamy v. Union of India* (2017) 10 SCC 1.

38 Section 3(7), (Draft) Personal Data Protection Bill, 2019.

39 Ryne Hager, "Google Confirms New Voice Confirmation Feature for Purchases in Assistant", *Android Police*, 25 May 2020, https://www.androidpolice.com/2020/05/25/google-assistant-gets-new-confirm-with-voice-match-setting-for-payments/.

40 Digidentity. "Privacy or Security? 'Function Creep' Kills Your Privacy", retrieved October 17, 2020, https://www.digidentity.eu/en/article/Function-creep-kills-your-privacy/

becomes particularly relevant when dealing with an individual's personal information or the information of people whose consent has not been obtained, such as children or people who are excluded from going through the privacy policies due to accessibility reasons or old age.

5. Conclusion

While VIs provide not just convenience but also an easier way to navigate the internet for some people, concerns around privacy and data protection loom large. While there is a need for VIs that are better at understanding the consumer, there is also a need to understand how these systems get their training data. With more voice technologies moving to always listening systems that can send targeted ads and use voice as a verification and identification system, there is a need to look closely at the privacy risks resulting from the collection, usage, and processing of voice data.