MAKING VOICES HEARD

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**
Copyediting **THE CLEAN COPY**
Illustration **KRUTHIKA N.S.**
Report Layout and Design **SAUMYAA NAIDU**

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)

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.

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 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]

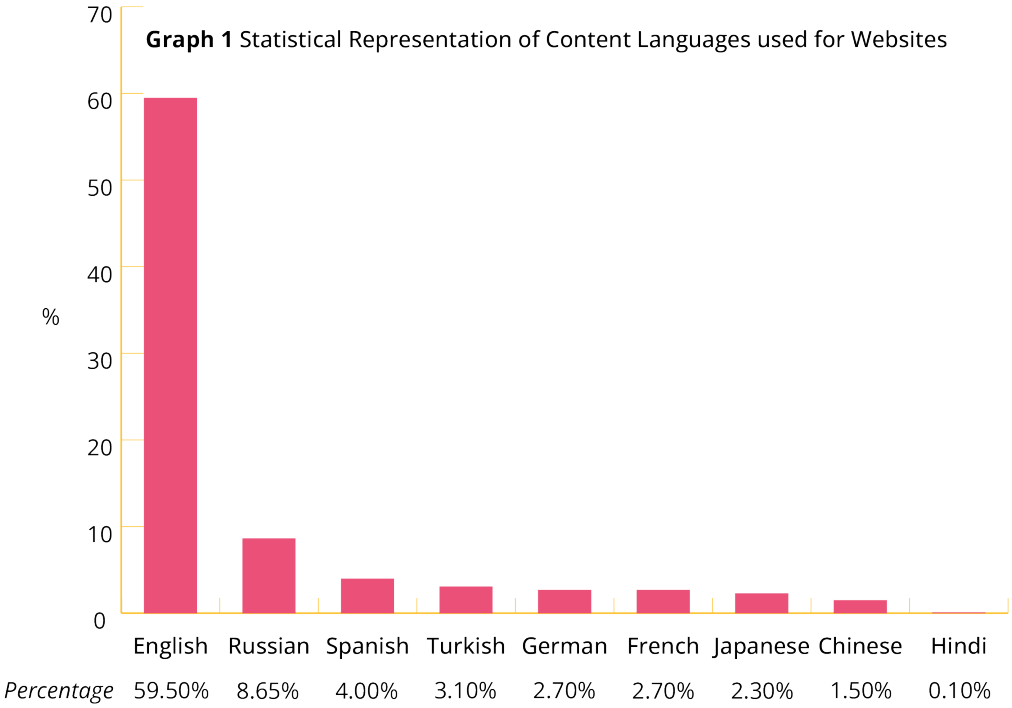
10 "Usage Statistics of Content Languages for Websites", *W³Techs*, accessed 3 November 2021 **https://w3techs.com/technologies/overview/content_language**.

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*.



Graph 1 Statistical Representation of Content Languages used for Websites

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

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.

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 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]

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**.

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.

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.

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]

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/.

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".

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, 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

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.

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*.

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 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

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*.

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/**

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 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

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.

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*.

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.

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*.