Kristu Jayanti Journal of Computational Sciences Volume 2 2022, Pages 57-70

Comprehensive Assessment of Automatic Speech Recognition System for building Artificial Intelligent Schemes

[1]Alex Joseph, [2]Dimal Thomas, [3]Merlin Reji

Department of Computer Science [PG]
Kristu Jayanti College (Autonomous), Bengaluru, India
[1,2,3]alexjoseph513.aj@gmail.com, dimalthomas10@gmail.com ,merlinreji7820@gmail.com

Abstract

Speech recognition is a multidisciplinary branch of natural language processing (NLP) that allows machines to recognize and translate spoken language into text. Speech recognition is crucial in the digital transformation process and it is widely employed in a variety of fields, such as education, industry, and healthcare, and has lately been employed in a number of Internet of Things (IoT) applications. Speech is a simple and effective method for human communication, but nowadays we are not only connected to one another, but also to the various devices in our lives. As a result, this kind of communication can be used by both computers and people. Interfaces are used to carry out the interaction, which is referred to as Human Computer Interaction (HCI). This paper provides an outline of the key areas of Automatic Speech Recognition (ASR), an important topic in artificial intelligence. It also includes an overview of recent major research in speech processing, as well as a basic idea of our proposal that may be considered as a major contribution to this field of research. The paper also refers to some specific enhancements that can add value to the researchers in future.

Keywords: - *Speech Recognition, Speech Understanding, ASR Automatic Speech Recognition Systems, Hybrid Systems, Low resource Languages.*

1. Introduction

Human beings communicate with one another using a variety of means, including voice, hand gestures, facial emotions, and so on. However, voice is the most significant mode of communication for humans because it enables conversation and is the most acceptable way of communication between people. Speech is a useful expression with a specific meaning that is made up of multiple words, each of which has several letters and voices. This voice can travel in the form of waves via air or any medium. Logical communication is delivered by speakers who speak the same language, implying that the sender and receiver share the same set of keys for understanding the message.

The researchers exploited this phenomenon and made a crucial branch in human-machine communication, where sound has aided in the user's usage of the computer and the creation of natural dialogue between them. Automatic voice recognition has aided the development of artificial intelligence, which aims to provide a very flexible means of manipulating machines, allowing users to converse and share information without the use of traditional input/output modules like the keyboard. Voice-based input/output systems are beneficial in a variety of situations, including the care of disabled individuals, the usage of automobiles, particularly while driving, distress calls, and so on.

2. Automatic Speech Recognition

Automatic voice recognition is one of the most automated aspects of speech processing, allowing the machine to interpret the user's speech and transform it into a series of words using a Machine Learning Algorithm, resulting in a natural conversation between humans and machines. Automatic voice recognition, commonly known as speech recognition, is a graphical depiction of emitted frequencies as a function of time. Voice interfaces and voice interaction can be created using any speech processing approach [2][3][6].

Voice Recognition can be used in a variety of situations, including:

- Voice services, such as a talking clock, weather forecasts, racing results, and so on.
- Entry of data and controlling the quality.
- Avionics and Flight Training.
- Vocal Dictation for the disabled.

Also, embedded voice recognition modules are found in mobile phones and automobiles, such as car radios, air conditioning, and onboard navigation on the Internet utilizing the voice commands [3][6][7].

3. ASR Evolution

Since the middle of the twentieth century, computer scientists have been trying to figure out how to make computers and humans understand each other. Speech-recognition technology has come a long way since its humble beginnings in the 1950s, with the first speech recognizer constructed in the 1950s and the voice assistants that are used on a daily basis.

3.1. Quick breakdown of the history of ASR

In 1952, HK Davis developed Audrey at Bell Labs, an autonomous numeric recognition computer. It can pronounce digits from 0 to 9 with an accuracy of over 90%. It also worked great with other speakers with 70-80 percentage accuracy [20]. IBM launched its "Shoebox" machine. The Shoebox could understand spoken numbers 0 to 9, as well as terms like "subtotal”,” minus”,” plus”, “false”, “total”, and “off”. Researchers in the Soviet Union in the late 1960s developed a dynamic temporal warping method that allowed recognizers to understand about 200 words [14].

In 1971, there was a huge growth in terms of voice recognition technology. The US Department of Defence funded for Voice Understanding Research (SUR) for a speech recognizer that can understand thousands of words. A speech-recognition system named Harpy can understand 1011 words [20]. In the ‘80s Fred Jelinek created Tangora in collaboration with IBM. Tangora is a speech recognition typewriter with a vocabulary of 20,000 words. It uses some specific parameters in recognizers and it can predict the speech pattern and data using a statistical approach [14].

In 1997, Dragon System developed continuous dictation software named “The Dragon Naturally Speaking”. It recognises only one word at a time, with a recognition rate of 100 words per minute, and it is feasible for speech-to text applications [20]. In the 2000s, by using Machine Learning Algorithms several updates happened in the field of automatic speech recognition in terms of accents, pronunciation, and context. In the year 2008, The Google Mobile App is

launched. It provides voice search for iPhone mobile users. Later, Google introduced a personalised voice recognition on Android. In 2011, Apple enhanced its voice search, which is coined as Siri and Microsoft introduced its own voice assistance called Cortana [14].

4. How ASR works

The working of the ASR system starts when the computer gets the audio from the source. As it receives the audio, it is divided into small units of speech and then this small component of speech is identified by the Automatic Speech Recognition System in the computer and transformed it into printed format. Some ASR systems are speaker-dependent, require extra training to identify the words and patterns. So that the ASR system can identify your voice, pattern and accent. And this process is done by the Machine Learning Algorithms. So that the ASR system can improve itself. But there are ASR systems that require no training and speak-independent systems [13][12].

ASR systems are dependent on the presence of a speaker and are used for interactive voice requests and responses like Siri, Google assistance, Alexa and so on. The ASR system is basically divided into three major components which convert the speech into its text format. They are Lexicon, Acoustic model, Language model [11][9][8].

4.1. Basic Steps:

- You use an audio feed to communicate with the software.
- The device you're speaking to records your speech as a wave file.
- Background noise is removed from the wave file, and the loudness is normalised.
- The filtered waveform is then split down into what are known as phonemes. Phonemes are the fundamental sounds that make up language and words. There are 44 of them in English, and they are made up of sound blocks like "wh", "th", "ka", and "t".
- Each phoneme is like a chain link, and the ASR programme employs statistical probability analysis to deduce whole words and eventually complete sentences by examining them in order, starting with the initial phoneme.
- Your ASR can now respond to you in a meaningful way, having "comprehended" your words.

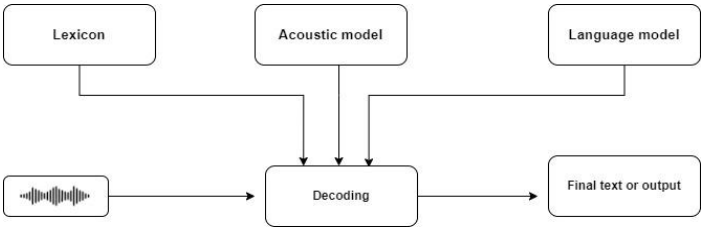


Fig 1 Working of ASR

The above given figure (Fig 1) is the diagrammatic representation of the basic steps of conversion of an audio file to a meaningful text. And The decoding is done by using three models, a brief description about those models is given below.

4.2. Lexicon

Lexicon is the first and initial stage in voice decoding. In an ASR system, a complete lexical design means that it should have the characteristics of both spoken language and written vocabulary (Input and Output). A word can be pronounced differently and, accordingly the meaning changes. The word "read," is pronounced differently depending on whether it is in the present or past tense in English. To ensure accuracy, comprehensive lexicons always check for all kind of pronunciation of a specific word. ASR systems employ lexicons that are tailored to each language. The lexicon is the starting step or the first input for the acoustic models for every voice input [2][3].

4.3. Acoustic model

In Acoustic modelling the inputted audio frame is divided into discrete time frames. Acoustic models examine each frame and predict the likelihood of smallest unit of the sound that is being used in that segment of audio. It is used to predict which sound will be spoken in each frame [1][12]. In a conversation, people say the same sentence in a variety of ways. Depending on the speaker, background noise and accents can make the same speech seem different. By using deep-learning algorithms that are being trained for several hours of diverse audio recordings and the potentially relevant texts Acoustic models assess the link between audio frames and smallest unit of sound [16][17][18].

4.4. Language model

Natural language processing (NLP) has been used to know or understand what a speaker says. The main duty of language models is to understand the meaning of spoken sentences and use this understanding to build word sequences. Similar to auditory models, they employ trained deep neural networks to predict the likelihood of which word will come next in a sentence. In the conversion of spoken languages to texts the most commonly used language model is N-gram probability [2][3].

E.g.: A string of words is known as an N-gram. "Contact centre", is a 2-gram term, while "Omni-channel contact centre" is a 3-gram term. Based on known prior words and basic grammar rules, N-gram probability predicts the next word in a series.

The accuracy of an ASR system is determined by calculating WER. WER can be calculated by using the following formula:

WER = substitutions + insertions + deletions / the number of words spoken

Factors including a speaker's pronunciation of specific words, recording or microphone quality, and background noise can all impact the WER of a speech-recognition computer. Even if the aforementioned defects are present, the user may find the decoded audio input helpful in a variety of situations. It is always essential to know that the usability of speech recognition software should not be evaluated exclusively by it [2][3].

5. ‘‘Tuning Test’’-How ASR is made to “Learn” from Humans

5.1. Human Tuning

ASR training may be completed in this manner in a relatively simple manner. It demands human programmers scan the conversation logs of a certain ASR software interface for often used terms that it needed to hear but didn't have in its pre-programmed lexicon. These words are then sent into the algorithms, allowing them to enhance their voice recognition [19][22].

5.2. Active learning

Active learning is a more advanced variant of ASR that is being tried alongside NLP voice recognition technology. The software is built to learn, remember, and adopt new words on its own via active learning, allowing it to constantly expand its vocabulary as it is exposed to new ways of speaking and expressing things. This helps the programme to pick up on a user's more specialised speech habits and engage with them more successfully, at least indirectly. If a human user consistently rejects autocorrect on a word, the NLP software learns to recognise that person's unique use of the word as the "correct" form over time [19][22].

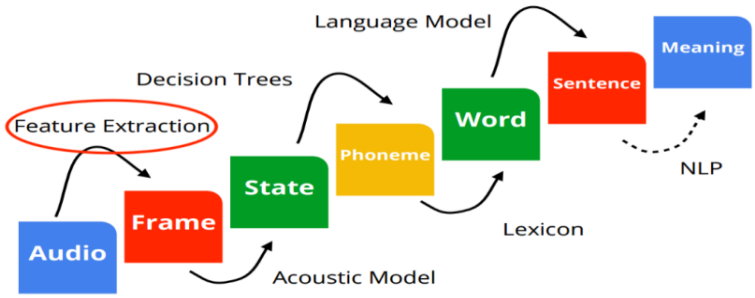


Fig 2 ASR Learning

The above given figure (Fig 2) gives an idea about how ASR system is made to learn from humans. And the stages of conversion of an audio file to a meaningful text is been plotted in the above figure. The three models Lexicon, Acoustic model, Language model are used for the decoding of an audio, and the stages inside these models is also plotted.

6. ASR Variants

6.1. Directed Dialogue

At work, Directed Dialogue dialogues are a much simpler variant of ASR, consisting of machine interfaces that direct you to respond verbally with a single word from a restricted set of alternatives, thereby crafting their answer to your carefully described request. In automated telephone banking and other customer service interfaces, directed dialogue ASR software is widely used [3][8].

6.2. Natural Language Conversations

Natural Language Conversations are more sophisticated versions of ASR that aim to mimic actual conversation by allowing you to communicate with them in an open-ended fashion rather than using a limited vocabulary. The Siri interface on the iPhone is one of the most advanced examples of this technology [3][8].

7. Challenges Faced by ASR

7.1. Accuracy

Accuracy nowadays refers to a lot more than simply the word output – the WER. On a case-by-case basis, many other factors influence the amount of accuracy. These elements are frequently specific to a use case or a certain business requirement, and they are Background noise, Punctuation placement, Capitalization, Correct formatting, Timing of words, Domain-specific terminology, and Speaker identification [15].

7.2. Data Security and Privacy

According to the Speechmatics poll, the concerns about privacy of ASR increased from 5% to 42%. Actually, the people's concerns were increased by the media's portrayal of internet giants as "data-hungry" [15].

7.3. Deployment

ASR technologies or software's installation should be always simple and clear and also the integration must be straightforward and secure, regardless of whether a company needs on-premises, cloud, or embedded implementation. If we don't have proper awareness, then the integration becomes very difficult. To overcome this barrier to acceptance, technology vendors must make installations and integrations as simple as possible [15].

7.4. Language Coverage

Several of the main speech technology companies have gaps in their language coverage. Although most providers provide English, when multinational organisations seek to use speech technology, a lack of language coverage is a barrier to adoption. Even when more languages are available, proper accent or vocabulary identification is frequently an issue [7][2][9][11].

7.5. The Voice Privacy Problems

Now voice technology has grown to its peak. Hence, the companies started to store the voice data for easy retrieval and study. This increased the privacy concerns. And the collected data is stored in the cloud.

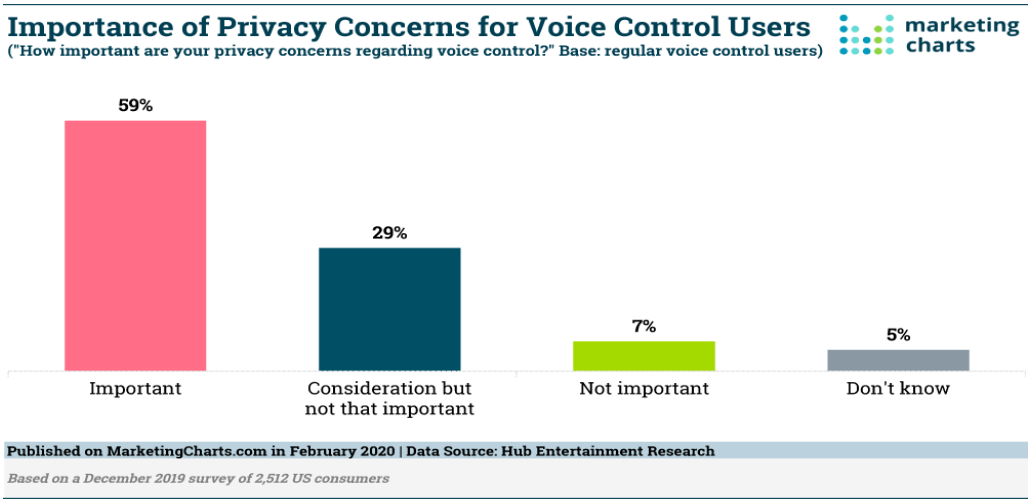


Fig 3, Image-Marketing Charts (Published on marketingCharts.com in Feb 2020|Data source: Hub Entertainment Research)

According to research conducted by Hub entertainment research globally, 45% of people are concerned about voice data privacy and 42% are concerned about voice hacking. In the case of voice control gadgets, 59 percent of respondents stated privacy is vital [Fig 3].

General Data Privacy Regulations classify voice recordings as personal data and subject to protection. When the companies are dealing with voice data, The European Digital Radio Alliance and the Association of European Radios made an agreement to apply the Digital Markets Act regulation to voice assistants [15].

7.6. Risks of Voice Technology Uses

Consumers and different organisations transfer a lot of information by using voice assistants and other voice related devices. And this has been stored in the cloud to process and transmit back the replies. And also, this data has been studied and a new algorithm has been made to increase the accuracy. The data that is being stored can be both very sensitive and also have common things [15].

Data storage in the cloud is costly, and the transmission of data from local apps and devices to the cloud and back might take time. And as a result, there is a chance for making the programmes and algorithms more complicated, which will give an advantage to the hackers for easy access. Once the cybercriminals get access then they can view the important information and use the information for biometric identification [15].

7.7. Consumers Voice Privacy Issues

Voice data can be used as biometric data for a person. When this data is stored locally, there is no much issue but when it is stored in the cloud, some abuses can occur. Thousands of complaints have been filed against tech giants like Google, Amazon, and Apple for improperly capturing and analysing voice recordings for targeted advertising or software enhancement, which is actually a violation of the privacy policy laws. General Data Protection Regulation is actually broken here [15].

7.8. The Threat of Insecure Speech Data to Enterprises

Enterprise data, especially private data, is a privacy and security priority. As a result, it is critical to ensure that competitors or other unwelcome parties do not have access to key company information. During the COVID-19 period, all the cooperating started to conduct meetings online by using several video conferencing applications, and these meeting recordings are actually saved and the misuse of these data can actually happen. Zoom was the most commonly used software. When several new users came and accessed the software at the same time lots of bugs occurred. Zoom is also having customer's video and audio recorded despite claiming to offer end-to-end encryption [15].

Zoom uses TLS encryption. Anyone attempting to access your information will be unable to hear or see your audio or video, but the Zoom organisation is still able to view the encrypted things. People use many smart voice assistants. For example, include an always-on microphone, allowing for unintentional recordings to be transferred to the cloud. The certain policies which are been proposed will damage the business. The main focus of a company is to safeguard traditional data, but speech data is also important. Fines of up to 20 million euros ($23 million US) can be imposed if the GDPR laws are broken [15].

7.9. Is the Cloud Secure Enough?

Cybercriminals can attack cloud computing platforms in a variety of ways. Data breaches, insider threats, and account hijacking are some of the most prominent cyber security dangers to cloud computing. Through this, Cybercriminals can view edit and transfer data including voice data. 60% of data breaches are actually done by IT users, Managers etc. The primary reasons for insider attacks are fraud and monetary gain, other threats to cloud-stored data originate from human error or ignorance [15].

Cybercriminals can access Cloud Computing by staff account, password cracking, or phishing emails are external dangers to company data. Cyber security is the most significant difficulty that businesses face in order to stay up with the demand for technology. Voice is a uniquely human trait that could be exploited to divulge personal information, conduct business, or perpetrate identity fraud [15].

7.10. Solving the Voice Privacy Problems

To avoid voice spoofing, businesses should adopt multifactor authentication. As a backup for identity verification, another biometric can be employed. In case of really sensitive information businesses should employ a variety of verification procedures The Voice Privacy Alliance's principles should be followed to protect voice data [17].

7.11. Voice AI comes down from Cloud to the Edge

An alternative way to store voice data is to store the data in edge. The term "edge" refers to data processing that takes place closer to the devices that gather it rather than sending it to a central point. It eliminates latency difficulties with cloud-processed data and also makes the data more secure. Google has been discreetly collaborating with local AI to improve the performance of neural networks on IoT devices. Amazon has taken it a step further with its latest Echo products. Instead of uploading recordings to the internet, its smart speakers and displays allow

users to record voice commands locally. The start-up claims to be the first to provide smart speakers with this privacy-first option [23][24] [25].

Kardome has created a spatial voice technology that improves speech recognition, recognises the speaker, and records the user's positions on edge devices. This technology will allow businesses and individuals to use voice-enabled devices without having to connect to the cloud. All voice data is handled on-site and is kept private [24].

The below given figure (Fig 4) explains how the audio files is been stored in edge instead of regular cloud storages, and also it explains how these audio files is been retrieved from edge and through what medium we can access these files. The platform where we will be able to retrieve and use the audio files is also listed [23] [25].

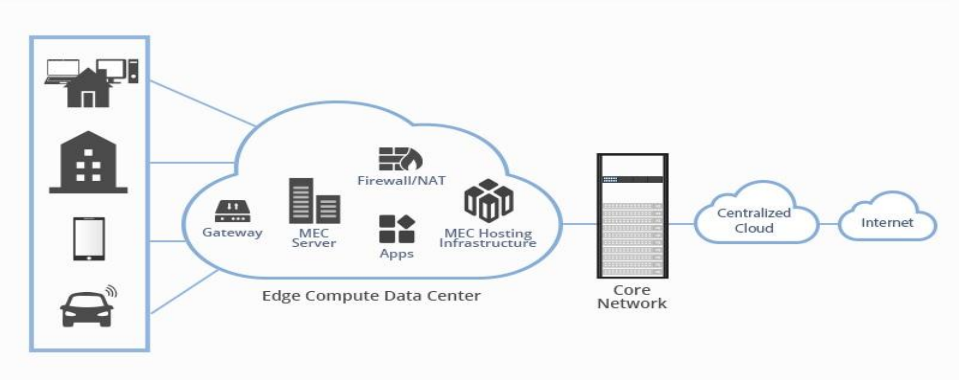


Fig 4 Edge Computing [12]

8. Future of ASR

- Mobile App Integration
- Voice-Tech in Healthcare
- Search Behaviors Will Change
- Individualized Experiences
- Voice Cloning
- Smart Displays
- Voice in gaming Industry

9. Results and Discussions

Table 1 shows the comparison results based upon a questionnaire answered by a population of 200 people and Figures from Fig 5 to Fig 14 is exactly the graphical representation of the below given table which was obtained from a population of 200 people. From the given table (Table 1) and graphs (Fig 5 to Fig 14) we can find the Average age of the ASR user's, which is the most commonly used ASR System, The satisfaction level of the users, Accuracy rate and efficiency

rate of the embedded ASR Systems, Privacy exploitation, Basic purpose, Policies, Day to day life and ASR.

Table 1: Comparison of Parameters

Average age of the ASR System user's	Below 10 0.15 %	10-30 70 %	30-40 14 %	40 + 15.75 %
Most commonly used ASR System	Siri 13 %	Alexa 15 %	Bixby 0.5 %	Google Assistance 71.5 %
User's satisfaction level	Good 62.5 %	Fair 32 %	Poor 7.5 %	--
Accuracy rate of ASR System's	Good 45.5 %	Fair 47.5 %	Poor 7.0 %	--
Efficiency rate of Embedded ASR	Good 48.5 %	Fair 41.2 %	Poor 10.1 %	--
Does Embedded ASR Exploit Privacy	Yes 15.6 %	No 44.2 %	Not-Sure 40.2 %	--
Makes your day-to-day activities easier	Yes 46.2 %	No 13.6 %	Maybe 40.2 %	--
Does it support your local language	Yes 38.9 %	No 32.8 %	Maybe 28.3 %	--

Your Primary purpose of Embedded ASR	Searching 80.8 %	Translation 18.2 %	Put alarm 0.8 %	Navigate 1.0 %
Are you against the policies of Embedded ASR	Yes 23 %	No 40 %	Maybe 37 %	--

These graphs are being plotted by using the above table data which is been obtained from a questionnaire which was circulated among a population of 200 people.

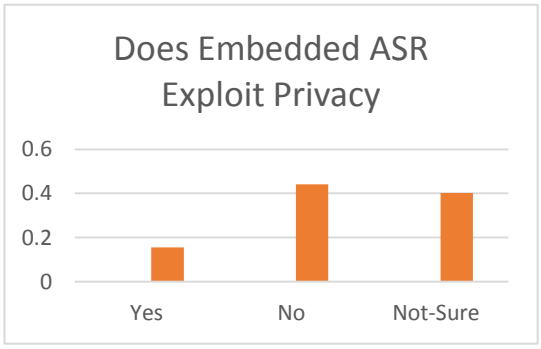


Fig 5: Privacy exploitation

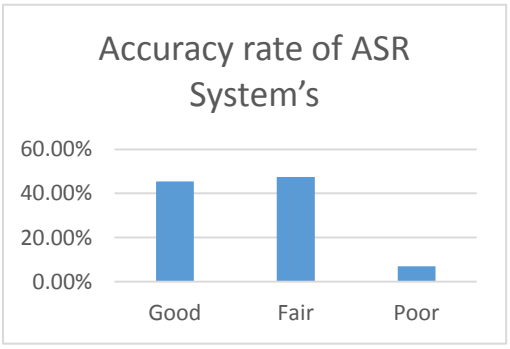


Fig 6: Accuracy rate of ASR

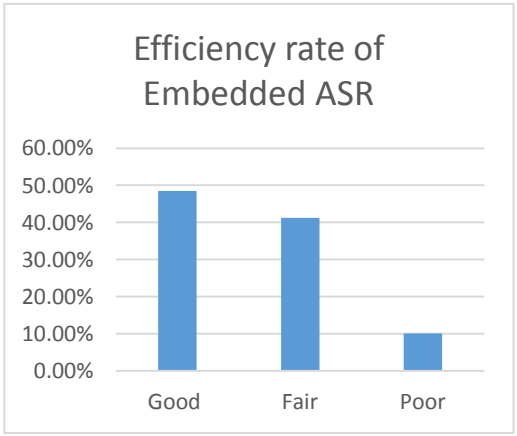


Fig 7: Efficiency rate of embedded ASR

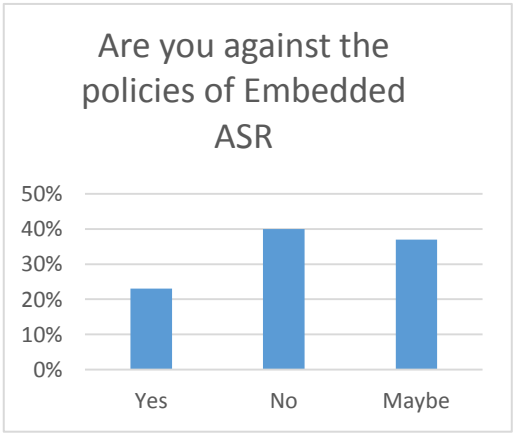


Fig 8: Against the policies of Embedded ASR

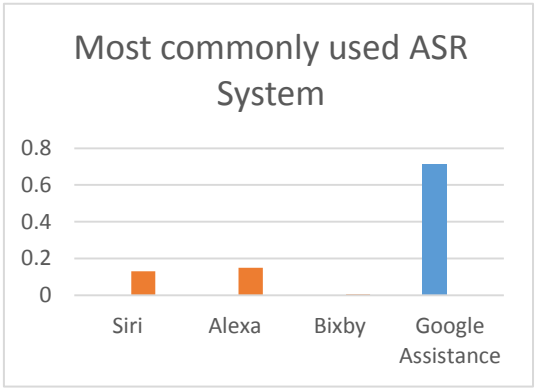


Fig 9: Commonly used ASR

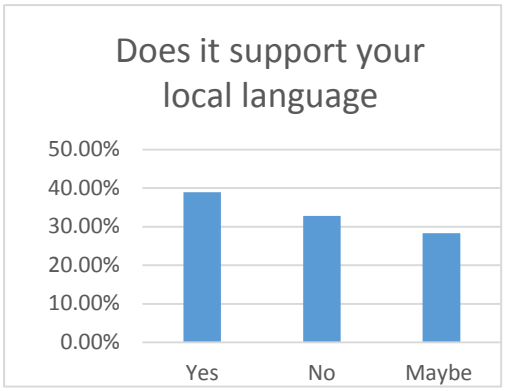


Fig 10: Support Local Language

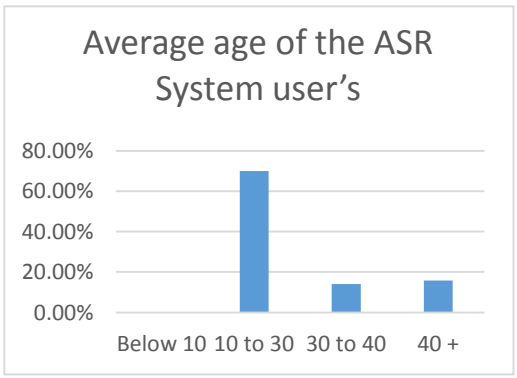


Fig 11: Average age of ASR System Users

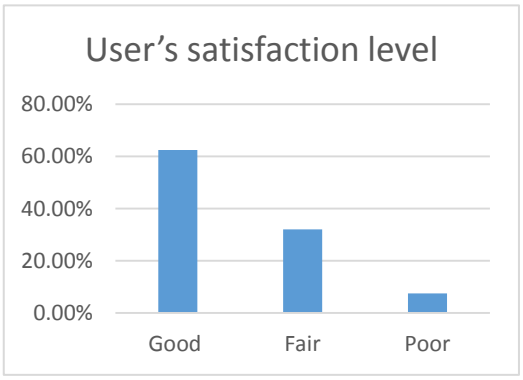
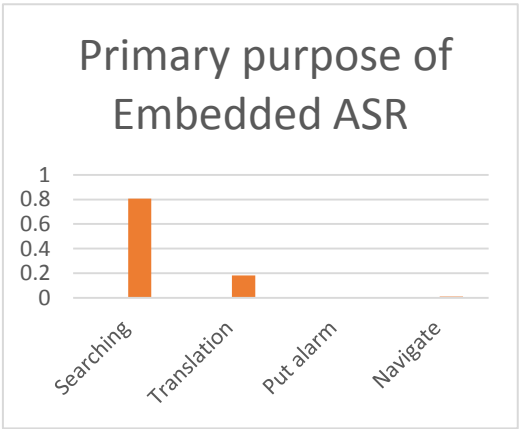


Fig 12: User Satisfaction level



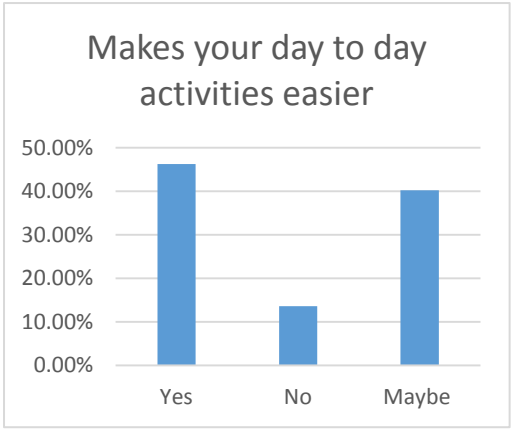


Fig 13: Primary purpose of embedded ASR

Fig 14: Daily life experience

Conclusion

It's difficult to get machines to listen. Even if we simply analyze current use cases and overlook the tremendous prospects it will offer, it's surprisingly strong. However, we must keep in mind that enormous power comes with great responsibility. As technologists, we must protect our users' privacy, design technologies that are free of bias and prejudice, and create systems that benefit everyone. Voice-based technology will continue to give more customized experiences as their ability to recognize and distinguish users' voices improves. However, the threat to voice privacy persists. On-the-edge by calculating speech instructions locally, Speech-AI addresses any identification and other privacy issues. Service providers may avoid privacy and compliance problems by avoiding sending customers' voices to the cloud, which fraudsters might use as a personal identity. The future work integrate neural networks into our approach to automatic speech processing to construct an intelligent interface based on computer vision that receives user voice commands, allowing intelligent interaction with users and establishing a natural and simple connection between the machine and the human. The user interface can also merge voice with other human senses and make decisions in real time.

References

[1] Hinton G, Deng L, Yu D, Dahl G.E, Mohamed A.R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T.N: Deep neural networks for acoustic model in speech recognition. IEEE Signal Process. Mag. (2012),29(6), 82–97

[2] Huang X, Acero A, Hon H.W: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, Englewood Cliffs (2001)

[3] Huang X, Acero A, Hon H.W: Spoken Language Processing, vol. 18. Prentice Hall, Englewood Cliffs (2001)

[4] Talluri Raj: Why edge computing is critical for the IoT. Network World (2017).

[5] Juang, B.H, Hou W Lee: Minimum classification error rate methods for speech recognition. IEEE Trans. Speech Audio Process., (1997),5(3), 257–265.

[6] Rabiner L, Juang B.H: Fundamentals of Speech Recognition. Prentice-Hall, Upper Saddle River (1993)

[7] S. Aalburg and H. Hoege: Foreign-accented speaker independent speech recognition. In Proc. of ICSLP, (2004) pages 1465–1468.

[8] M. Adda-Decker and L. Lamel: Pronunciation variants across system configuration, language and speaking style. Speech Communication, (1999),29(2):83–98,

[9] L.M. Arslan and J.H.L. Hansen: Language accent classification in American English. Speech Communication, ,(1996),18(4):353–367.

[10] B. Atal and L. Rabiner: A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. In IEEE Trans. on Acoustics, Speech, and Signal Processing, (1976),volume 24, 3, pages 201–212

[11] Mats Blomberg: Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references. Speech Communication, (1991),10(5-6):453–461.
[12] M. J. F. Gales: Acoustic factorization. Madona di Campiglio, Italy, 2001.
[13] P. L. Garvin and P. Ladefoged: Speaker identification and message identification in speech recognition. Phonetica, 1963.9:193–199.
[14] O'Shaughnessy D: Automatic speech recognition: history, methods and challenges, Pattern Recognit, , (2008),41 (10), pp. 2965–2979.
[15] Vimala C, Radha V: A review on speech recognition challenges and approaches, World Computer Science Information Technology, (2012),2(1), pp. 1–7.
[16] S. F. Boll: Suppression of Acoustic Noise in Speech Using Spectral Subtraction, IEEE Trans. Acoustics, Speech and Signal Processing, 1979,Vol. 27, pp. 113-120.
[17] M. Berouti, R. Schwartz and J. Makhoul: Enhancement of Speech Corrupted by Acoustic Noise, in Speech Enhancement, J. S. Prentice Hall, 1983, Englewood Cliffs, NJ.
[18] R. Stern and A. Acero: Acoustical Preprocessing for Automatic Speech Recognition, DARPA Speech and Natural Language Workshop, 1989.
[19] Machowski Michael: Speech Recognition and Natural Language Processing as Highly Effective Means of Human-Computer Interaction. University of Colorado, Department of Computing Sciences, 1997.
[20] McAllister, Alex: Voice/Speech Recognition Technologies Report and Tutorial. Bell Atlantic. May 23, 1995.
[21] Peacocke Recihar D, Graf Daryl H: An Introduction to Speech and Speaker Recognition. IEEE Computer (1990),23(8), pp 26 - 33.
[22] White George M: Natural Language Understanding and Speech Recognition. Communications of the ACM, (1990)Vol. 33, No. 8.
[23] Leksandrova, Mary: The Impact of Edge Computing on IoT, The Main Benefits and Real-Life Use Cases. Eastern Peak (2019).
[24] Nelson Patrick: How edge computing can help secure the IoT. Network World (2019).
[25] Caulfield Matt: Edge Computing, 9 Killer Use Cases for Now & the Future. Medium (2018).