International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 16, Issue.3,July 2023

CHALLENGES IN AUTOMATIC SPEECH RECOGNITION IN VOICE ASSISTANTS AND HOW TO MITIGATE THEM

Ashlesha Vishnu Kadam

Abstract: *Automatic Speech Recognition (ASR) has increasingly been finding more and more applications, especially because of the growth seen in smart speaker awareness, adoption and engagement, and ASR being a critical early step in the end-to-end voice assistant usage. However, despite advancements, there continue to be gaps in ASR. This article explores those gaps and provides strategies to overcome those challenges. The gaps and challenges have been articulated specifically from the point of view of the use case of music and music-related queries on voice assistants.*

Keywords*: ASR, NLP, TTS, voice assistant*

I. INTRODUCTION

Automatic Speech Recognition, also known as ASR, or STT (speech to text), is the term used for transcribing spoken language into text using technology [1]. ASR is a common first step in enabling interactions between humans and machines. Consciously, or unconsciously, you might have used this technology many times. For example, while using the mic icon in your Google search bar, when talking to Siri while driving a car, when asking Alexa to play music or when "talking" to a customer service system.

While ASR has made strides over the last few years, it is far from smooth, intuitive and glitch-free. Think about the number of times you had to correct yourself or repeat what you said, or change your choice of words or volume when talking to Siri, Google Assistant or Alexa. There are many challenges even today when it comes to ASR. Let's look at what some of these challenges are, and if there are ways in which these can be mitigated in order to provide an improved user experience on voice assistants.

II. HOW ASR WORKS

Before getting into understanding the challenges with ASR, it is helpful to understand how ASR works. Let's use a real-life example of a user asking a hypothetical voice assistant Nova for a music track.

As soon as the user says something like "Hey Nova, can you play the song Fallin'?" expecting the song by artist Alicia Keys, the voice assistant Nova receives this input from its mic. The audio file is then pre-processed so as to remove any ambient noise, boost the signal and extract features from this audio file that are relevant for speech recognition [2]. This audio input is then fed into a language model that has been trained on large amounts of speech data [4]. This model is able to determine the likeliest string of tokens that match the audio file. The output of this model is this string of tokens. In our example, this would mean that the model's output is a text output similar to: "can", "you", "play", "the", "song", "fallin'", assuming no errors. See Fig. 1. for an overview of the ASR process.

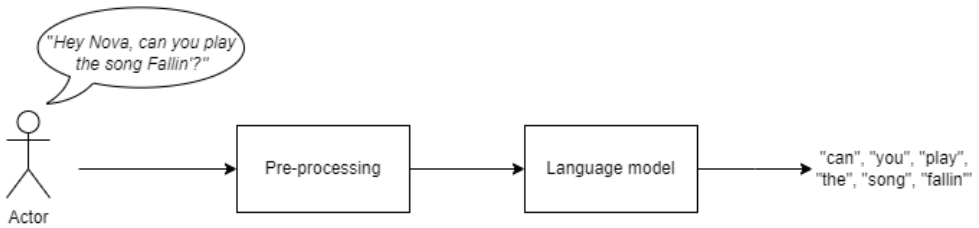

Fig. 1. Overview of ASR process

Technical Product Lead, Amazon Music, Amazon.com LLC
2021 7th Ave Seattle 98121 WA USA

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 16, Issue.3,July 2023

While this is a simplified overview, ASR technology is highly complex, relying on advanced algorithms and statistical models to accurately transcribe spoken language into text. However, most ASR implementations don't seem to work as smoothly. Let's look into what some of the challenges in ASR are that result in a suboptimal experience.

III. CHALLENGES IN ASR

Following are some of the key challenges in ASR in the context of voice assistants:

3.1 Different Speaking Styles

When considering a global user base, different customers use different languages, making ASR a challenge. Even within the same language, there might be different accents, dialects, enunciations, speaking styles and personal idiosyncrasies [5][6]. For example, the UK along has over 40 different dialects [7] that sound completely different from each other and often use different spellings and word structure. These variations in speech makes it particularly challenging for ASR systems to correctly transcribe speech to text.

3.2 Ambient noise

In addition to the variability in speech mentioned above, ambient and background noise can also interfere with ASR [8]. For example, Nova might perfectly understand "*play Lady Gaga*" when in a quiet setting, like your work den. But the same request might be understood as "*play Radio Ga Ga*" while driving with noisy passengers. Other similar situations could be when there's a party with friends and everyone's talking, or when there is some construction happening right outside your house, and so on. Handling these unexpected sources of noise and interference is challenging for ASR technology [9].

3.3 Unfamiliar Words and Phrases

Users might use words that are seldom used, including slang, colloquial terms, specific jargon, aliases or even words that are made-up (i.e. not part of the standard dictionary). Since these words might appear sparsely in training data too [10], ASR systems commonly fail to correctly recognize and transcribe these words to text [11]. For example, if you happen to know the name of the owner of a local grocery store, and ask Nova in the car to "*show the route for Tom's shop*", there's a high likelihood of Nova failing to transcribe what you said correctly. Other examples include words, or sequence of words that are not out of vocabulary but might simply be new to the world. For example, when an artist releases a new track, or if a track isn't very popular, the title of that track might be unfamiliar to a voice assistant and hard to transcribe. For instance, you might ask Nova to "*play Theodosia"* but Nova might not transcribe this correctly and start playing something else instead of the track from Hamilton.

3.4 Contextual requests

Sometimes users might say something that could have multiple interpretations. It is hard for ASR systems to disambiguate across these interpretations to find the most accurate one [12]. For example, if a user says something to Nova and because of the ambient noise, the only part that Nova gets is "*Nova, [..] washing machine*". It would be hard for Nova to understand if the user meant Nova should start the washing machine or play the song Washing Machine by Sonic Youth. Similarly, if ASR as a capability is planned to be used for a specific domain or set of domains, a generic ASR model might not do well [13]. An example of this is using a generic ASR model vs. a model trained specifically for music. A user request like "*Nova, play subtract*" might not make sense to a generic ASR model but there's a likelihood that it will be correctly understood by an ASR model trained on the music catalog as being a request for playing the artist Ed Sheeran's album, Subtract.

3.5 Personalization to Speaker

Users often run into situations where a voice assistant fails to understand what they asked for, even though it might be something that they have asked for many times in the past, which can be quite frustrating and sometimes even lead to confusion and lack of trust in the voice assistant [14]. For example, if you have asked Nova in the past to play music by the American Jazz musician David Gilmore, you assume Nova understands you like this band. So, if you say "*Nova, play David Gilmore*", but Nova starts playing music by the more popular David Gilmour of Pink Floyd, it might be confusing and lead to the feeling of "Nova still doesn't get me".

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 16, Issue.3,July 2023

IV. MITIGATIONS TO ADDRESS ASR CHALLENGES

While there are many challenges when it comes to ASR, there are some mitigation strategies that help address some or all these challenges to varying degrees. Some common techniques to overcome ASR challenges are mentioned below.

4.1 Diverse training data

The data that the ASR models are trained on needs to capture as much diversity as possible in terms of speech variability [15]. For example, training the ASR models on different ways in which someone might say something, as well as different users and how they'd say the same thing. For example, if you want to teach Nova to understand requests for live music, it would mean that Nova needs to be trained on a corpus of data where users with different languages as their primary language, users with different accents, styles of speaking, and speech idiosyncrasies have pronounced the word "live". Without this, Nova might constantly keep mis-transcribing the word as "life", "libe", "love", and so on.

4.2 Data augmentation

To mitigate the risk of variation in what customers ask for and how, data augmentation techniques like speed perturbation (i.e. deliberately introducing distortion in speech signals to make ASR models more robust), reverberation simulation (i.e. training the ASR models with synthetically created reverberant speech signals), noise injection (i.e. deliberately introducing sounds and noise to speech signals) and more should be used to enhance ASR models [16]. A nifty approach could be generating synthetic request traffic at scale, applying text to speech (TTS) technology to it, and then using this data for training or fine-tuning models can help.

4.3 Multi-lingual and sparse word handling

Families are increasingly becoming multi-lingual. For an application based on ASR to scale globally and gain wide adoption and engagement, it needs to be able to understand multiple languages [17]. This is especially important for the application of voice assistants that tend to be used in family setting (e.g. in the kitchen or family room where usually more than one person uses them). Using techniques like multi-lingual training and code-switching modeling, the voice assistant can be training to understand inputs in multiple languages [18]. For example, if multi-lingual training is implemented for Nova, Nova would understand consecutive requests correctly, like "*Hey Nova, play Everlong*", "*Arrêt, arrêt! Play the acoustic version*" means stop playing the song and play its acoustic version instead. Note that "arrêt" means "stop" in French.

4.4 Transfer Learning from Pre-trained models

Pre-trained models like Large Language Models (LLMs) can be used to enhance ASR. Using techniques like fine-tuning and feature extraction, ASR models' ability to handle complex inputs can be improved [19]. An example is domain specific models like music or healthcare related solution that uses ASR. Fine-tuning the ASR models for language commonly used in these domains using LLMs might enhance the ASR.

4.5 Feedback signal

In order for the ASR models of a voice assistant to keep improving continually, it needs to capture user-initiated corrections. For example, if a user says something and Nova plays Radio Ga Ga but the user stops the music and repeats to Nova to play Lady Gaga, that signal should be used to remember the correct transcription of the sound signals into words. The ASR system also needs to consider the feedback about errors. For example, if the user stops Nova from playing something, and repeats the request, the barge-in should be fed as a signal to the ASR system for learning. ASR models also need to be updated regularly based on evolving user needs [20]. For example, it should happen that the user regularly runs into the same ASR errors every day. The model needs to get updated to know that even if the user's voice isn't very clear, it gets the right ASR interpretation.

V. CONCLUDING REMARKS

In conclusion, the importance and adoption of ASR is only set to increase. However, ASR suffers from some challenges that could make for a poor user experience. By using some of the ASR enhancement techniques mentioned above, these challenges could be mitigated to a great extent to enhance ASR and enable wider adoption.

REFERENCES

[1]. https://paperswithcode.com/task/automatic-speech-recognition
[2]. S. Alharbi et al., "Automatic Speech Recognition: Systematic Literature Review," in IEEE Access, vol. 9, pp. 131858-131876, 2021, doi: 10.1109/ACCESS.2021.3112535.

[3]. Jinyu Li, ... Yifan Gong, in Robust Automatic Speech Recognition, 2016
[4]. Cutajar, Michelle & Gatt, E. & Grech, Ivan & Casha, Owen & Micallef, Joseph. (2013). Comparative study of automatic speech recognition techniques. Signal Processing, IET. 7. 25-46. 10.1049/iet-spr.2012.0151.
[5]. Chao Huang, Tao Chen, Eric Chang, "Accent Issues in Large Vocabulary Continuous Speech Recognition" International Journal of Speech Technology 7, 141–153, 2004
[6]. Alicia Beckford Wassink, Cady Gansen, Isabel Bartholomew, "Uneven success: automatic speech recognition and ethnicity-related dialects", Submitted to: Speech Communication submitted: May 2020, revised: March 2021, second revision: January 2022
[7]. https://greatbritishmag.co.uk/uk-culture/how-many-british-accents-are-there/
[8]. Rajnoha, Josef & Pollák, Petr. (2011). ASR systems in Noisy Environment: Analysis and Solutions for Increasing Noise Robustness. Radioengineering. 20.
[9]. Shrawankar Urmila, Thakre Vilas, "Adverse Conditions and ASR Techniques for Robust Speech User Interface", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011
[10]. Cutajar, Michelle & Gatt, E. & Grech, Ivan & Casha, Owen & Micallef, Joseph. (2013). Comparative study of automatic speech recognition techniques. Signal Processing, IET. 7. 25-46. 10.1049/iet-spr.2012.0151.
[11]. Ketabdar, Hamed & Hannemann, Mirko & Hermansky, Hynek. (2007). Detection of out-of-vocabulary words in posterior based ASR. 4. 1757-1760. 10.21437/Interspeech.2007-492.
[12]. Amitoj Singh, Navkiran Kaur, Vinay Kukreja, Virender Kadyan & Munish Kumar, "Computational intelligence in processing of speech acoustics: a survey", https://doi.org/10.1007/s40747-022-00665-1
[13]. Michelle Cutajar, Edward Gatt, Ivan Grech, Owen Casha, Joseph Micallef, "Comparative study of automatic speech recognition techniques", https://doi.org/10.1049/iet-spr.2012.0151
[14]. Jari Kolehmainen, Yile Gu, Aditya Gourav, Prashanth Gurunath Shivakumar, Ankur Gandhe, Ariya Rastrow, Ivan Bulyko, "Personalization for BERT-based Discriminative Speech Recognition Rescoring" https://arxiv.org/abs/2307.06832
[15]. Mortaza Doulaty Bashkand, "Methods for Addressing Data Diversity in Automatic Speech Recognition", Machine Intelligence for Natural Interfaces (MINI) Lab, Speech and Hearing (SPandH) Research Group, Department of Computer Science, University of Sheffield
[16]. Damania, Ronit Jitesh, Data Augmentation for Automatic Speech Recognition for Low Resource Languages, Rochester Institute of Technology ProQuest Dissertations Publishing, 2021. 28772334.
[17]. Mu Yang, Andros Tjandra, Chunxi Liu, David Zhang, Duc Le, Ozlem Kalinli, "Learning ASR pathways: A sparse multilingual ASR model", https://arxiv.org/abs/2209.05735
[18]. Rafal Černiavski, "Cross-lingual and Multilingual Automatic Speech Recognition for Scandinavian Languages", Uppsala University
[19]. Kunze, Julius, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier and Sebastian Stober. "Transfer Learning for Speech Recognition on a Budget." Rep4NLP@ACL (2017).
[20]. https://www.cloudskillsboost.google/focuses/13597?parent=catalog