Speech-to-Text Conversion

Introduction

Speech-to-text (STT) technology has become a familiar part of our daily lives, transforming spoken words into written text using artificial intelligence (AI). From virtual assistants for example Siri and Alexa to subtitles on YouTube, this technology is in everywhere. In this assignment, I am going to explain how STT works, some of the main challenges it faces, and how it’s used in real-world applications.

1. How Speech-to-Text Conversion Works

a. Audio Preprocessing

The first step in speech-to-text conversion is preparing the audio. This process is known as audio preprocessing, and it's all about cleaning up the sound so the system can focus on the speaker’s voice. It usually involves:

- Noise Reduction – Removing background noise like traffic, other people talking, or wind.
- Segmentation – Breaking up the audio into smaller chunks, such as individual words or sentences, to make it easier to analyse.
- Normalization – Making sure the audio is at a consistent volume and speed so the system can interpret it more accurately.

b. Feature Extraction

Once the audio is cleaned up, the next step is to extract important features from it. These features help the AI model understand the sound and identify speech patterns. Two common techniques used here are:

- MFCCs (Mel-frequency cepstral coefficients): This mimic how the human ear perceives sound, allowing the system to focus on speech-relevant tones.
- Spectrograms: These are visual representations of audio over time, showing frequency changes and helping the system detect patterns in speech.

c. Speech Modeling

Now that we have the essential features, machine learning models step in to figure out what’s being said. These models are trained on large datasets of spoken language and written transcripts. Some common models include:

- RNNs (Recurrent Neural Networks): These are good at processing sequences, making them ideal for speech data.
- LSTMs (Long Short-Term Memory networks): A more advanced form of RNN that can better remember context from earlier in the sentence.
- Transformers: These are modern and powerful models used by systems like Google Voice Search. They're faster and more accurate than older methods.

d. Decoding and Language Modeling

After the model has identified individual sounds, it needs to turn them into meaningful text. This is done through decoding, guided by a language model. Language models use context to predict what words are likely to come next. For example, if someone say "Can I have a cup of…", the system might predict "coffee" or "tea" based on common usage.

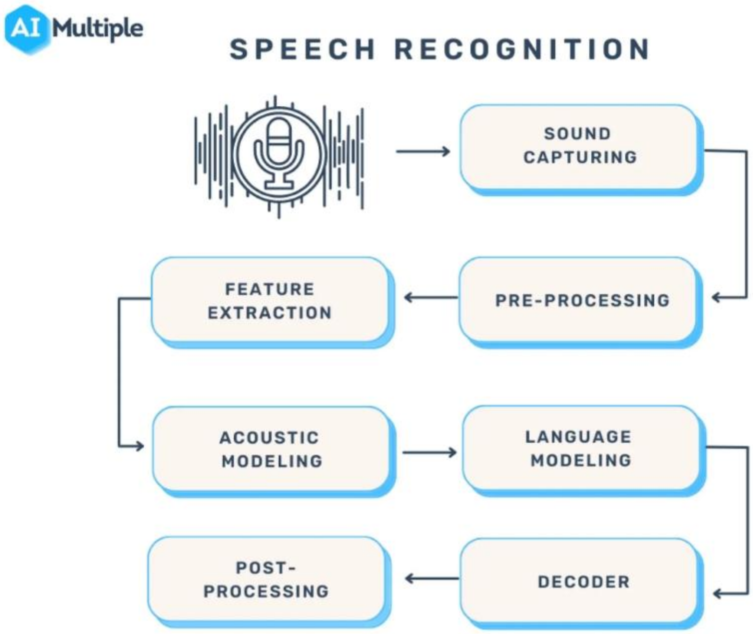

Figure 1Karatas, G. (2025) Speech recognition: Everything you need to know in 2025, AIMultiple. Available at: https://research.aimultiple.com/speech-recognition/ (Accessed: 24 March 2025).

The figure illustrates a speech recognition which is split down into several steps.

The process starts with the sound being captured, usually by a microphone, then it starts pre-processing where clean up the raw audio, removing background noise, normalize volume and move to feature extraction converting the audio to a machine language, after that use models to understand how sounds correspond to words and predict a better sequence of words using language modelling, combine information decoding and finally end the process.

2. Challenges in Speech-to-Text Systems

Despite all the advancements in STT technology, there are still a few challenges:

- Accents and Dialects: People from different regions pronounce words differently, which can confuse the system.
- Background Noise: Loud environments can interfere with accuracy, making it hard for the system to focus on the speaker.
- Multiple Speakers: When several people talk at once, the system may struggle to separate their voices or understand who's saying what.

3. Real-World Uses of STT

Speech-to-text is already integrated into many technologies we use every day:

- Virtual Assistants: Siri, Alexa, and Google Assistant rely on STT to understand user commands and respond appropriately.
- Automatic Subtitles: Platforms like YouTube, Zoom, and TV streaming services use STT to create subtitles, improving accessibility.
- Assistive Technology: For people with disabilities—especially those who are deaf, hard of hearing, or unable to type—STT is a vital communication tool.

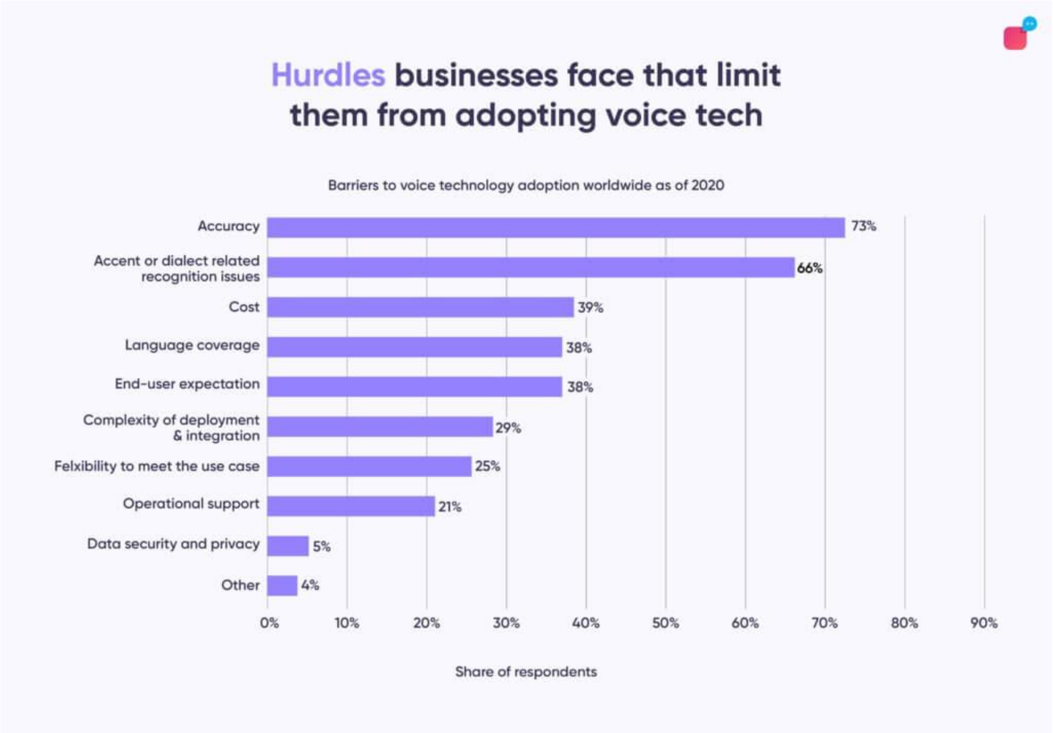

Figure 2 - Singh, G. (2022) 4 big challenges prevalent in speech Recognition Today, Verloop.io. Available at: https://www.verloop.io/blog/speech-recognition-challenges

As you can see in Figure 2 the challenges in speech-to text system using real world as an example.

The top Barries identified:

73% Accuracy - misinterpretation in spoken input, making bad to use on businesses.

63% Accent or dialect related recognition issues - Difficulty in understand different accents.

39% Cost - It is expensive to maintain.

38% Language Coverage - Limited support.

38% End-User Expectation - Any error leaves customers unhappy.

29% Complexity of Deployment and Integration – using existing voice tech system is challenging.

25% Flexibility to meet the use case - It doesn’t suit for all businesses.

21% Operational Support - inefficient IT support can cause difficulty in improvement.

5% Data Security and Privacy - Low concern but still relevant.

4% Other - less common concerns.

Conclusion

Speech-to-text is a remarkable example of how AI can bridge the gap between humans and machines. Through steps like audio preprocessing, feature extraction, speech modeling, and decoding, spoken language is turned into readable text. While there are still challenges, especially with accents and noisy environments, ongoing advancements in AI are making STT systems smarter and more accurate. As the technology evolves, we can expect to see even more useful applications in the future.

References:

- Jurafsky, D. & Martin, J.H. (2023). *Speech and Language Processing* (3rd ed.). Stanford University.
- O'Shaughnessy, D. (2008). *Speech Communications: Human and Machine*. IEEE Press.
- Google AI Blog. (2021). *Advances in Speech Recognition with Transformers*. Available at: https://ai.googleblog.com
- IBM Cloud Learn Hub. (2022). *What is Speech to Text?*. Available at: https://www.ibm.com/cloud/learn/speech-to-text