AIMultiple RESEARCH

Search topics

Speech Recognition: Everything You Need to Know in 2024

Gulbahar Karatas Voice recognition

Updated on **Jan 2** | 7 min read

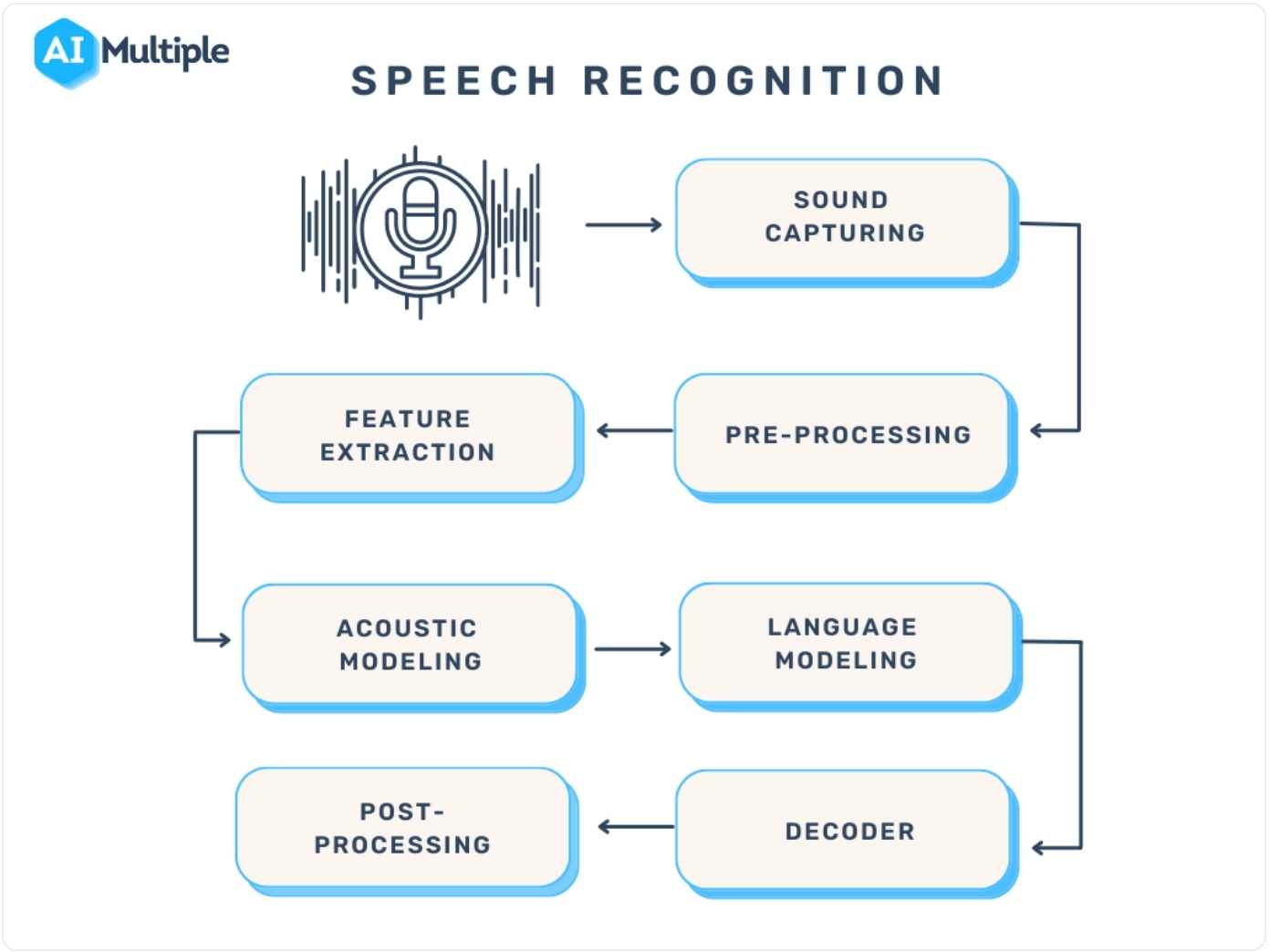


Table of contents

Speech recognition, also known as automatic speech recognition (ASR), enables seamless communication between humans and machines. This technology empowers organizations to transform human speech into written text. Speech recognition technology can revolutionize many business applications, including customer service, healthcare, finance and sales.

In this comprehensive guide, we will explain speech recognition, exploring how it works, the algorithms involved, and the use cases of various industries.

If you require training data for your speech recognition system, here is a guide to finding the right speech data collection services.

What is speech recognition?

Speech recognition, also known as automatic speech recognition (ASR), speech-to-text (STT), and computer speech recognition, is a technology that enables a computer to recognize and convert spoken language into text.

Speech recognition technology uses AI and machine learning models to accurately identify and transcribe different accents, dialects, and speech patterns.

What are the features of speech recognition systems?

Speech recognition systems have several components that work together to understand and process human speech. Key features of effective speech recognition are:

- **Audio preprocessing:** After you have obtained the raw audio signal from an input device, you need to preprocess it to improve the quality of the

speech input The main goal of audio preprocessing is to capture relevant speech data by removing any unwanted artifacts and reducing noise.

- **Feature extraction:** This stage converts the preprocessed audio signal into a more informative representation. This makes raw audio data more manageable for machine learning models in speech recognition systems.
- **Language model weighting:** Language weighting gives more weight to certain words and phrases, such as product references, in audio and voice signals. This makes those keywords more likely to be recognized in a subsequent speech by speech recognition systems.
- **Acoustic modeling**: It enables speech recognizers to capture and distinguish phonetic units within a speech signal. Acoustic models are trained on large datasets containing speech samples from a diverse set of speakers with different accents, speaking styles, and backgrounds.
- **Speaker labeling:** It enables speech recognition applications to determine the identities of multiple speakers in an audio recording. It assigns unique labels to each speaker in an audio recording, allowing the identification of which speaker was speaking at any given time.
- **Profanity filtering:** The process of removing offensive, inappropriate, or explicit words or phrases from audio data.

What are the different speech recognition algorithms?

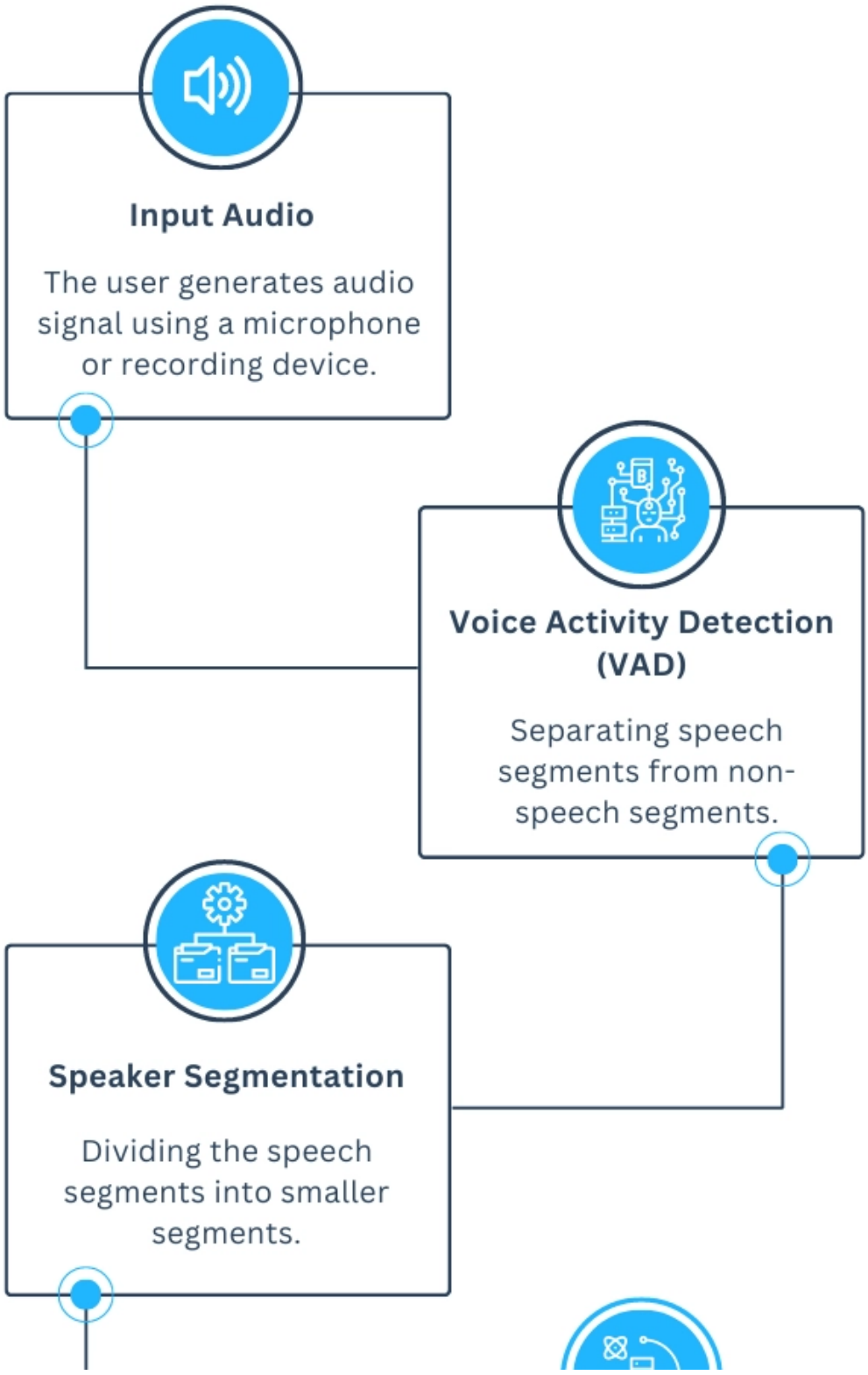
Speech recognition uses various algorithms and computation techniques to convert spoken language into written language. The following are some of the most commonly used speech recognition methods:

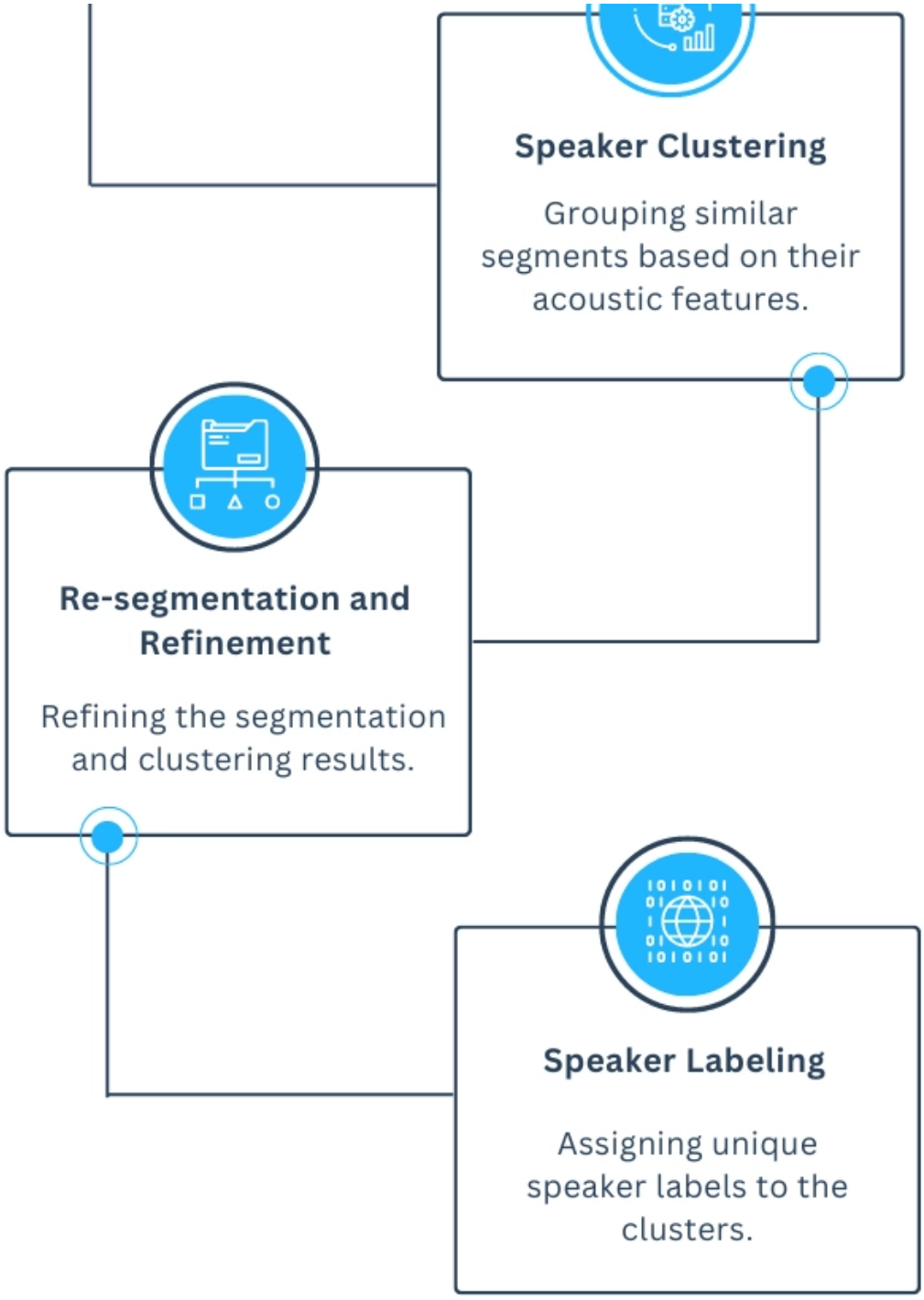
1. **Hidden Markov Models (HMMs):** Hidden Markov model is a statistical Markov model commonly used in traditional speech recognition systems. HMMs capture the relationship between the acoustic features and model the temporal dynamics of speech signals.
2. **Natural language processing (NLP):**NLP is a subfield of artificial intelligence that focuses on the interaction between humans and machines through natural language. Some of the key roles of NLP in speech recognition systems:
    - Estimate the probability of word sequences in the recognized text
    - Convert colloquial expressions and abbreviations in a spoken language into a standard written form

- Map phonetic units obtained from acoustic models to their corresponding words in the target language.

3. **Speaker Diarization (SD):** Speaker diarization, or speaker labeling, is the process of identifying and attributing speech segments to their respective speakers (Figure 1). It allows for speaker-specific voice recognition and the identification of individuals in a conversation.

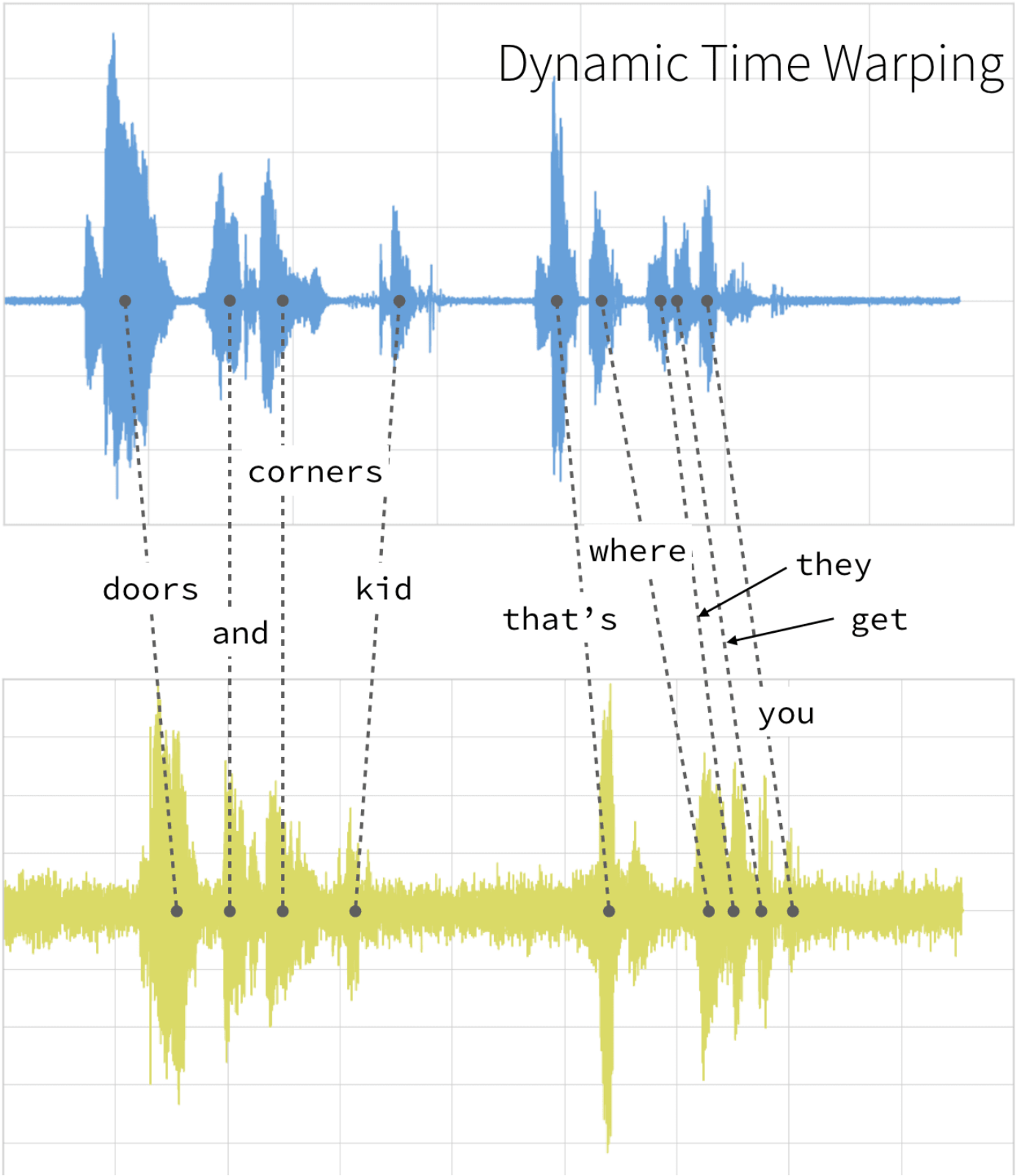
Figure 1: A flowchart illustrating the speaker diarization process


SPEAKER DIARIZATION PROCESS
Input Audio
The user generates audio signal using a microphone or recording device.
Voice Activity Detection (VAD)
Separating speech segments from non-speech segments.
Speaker Segmentation
Dividing the speech segments into smaller segments.



4. **Dynamic Time Warping (DTW):** Speech recognition algorithms use Dynamic Time Warping (DTW) algorithm to find an optimal alignment between two sequences (Figure 2).

Figure 2: A speech recognizer using dynamic time warping to determine the optimal distance between elements



Source: Databricks[1]

5. **Deep neural networks:** Neural networks process and transform input data by simulating the non-linear frequency perception of the human auditory system.

6. **Connectionist Temporal Classification (CTC):** It is a training objective introduced by Alex Graves in 2006. CTC is especially useful for sequence

labeling tasks and end-to-end speech recognition systems. It allows the neural network to discover the relationship between input frames and align input frames with output labels.

Speech recognition vs voice recognition

Speech recognition is commonly confused with voice recognition, yet, they refer to distinct concepts. Speech recognition converts spoken words into written text, focusing on identifying the words and sentences spoken by a user, regardless of the speaker's identity.

On the other hand, voice recognition is concerned with recognizing or verifying a speaker's voice, aiming to determine the identity of an unknown speaker rather than focusing on understanding the content of the speech.

What are the challenges of speech recognition with solutions?

While speech recognition technology offers many benefits, it still faces a number of challenges that need to be addressed. Some of the main limitations of speech recognition include:

Acoustic Challenges:

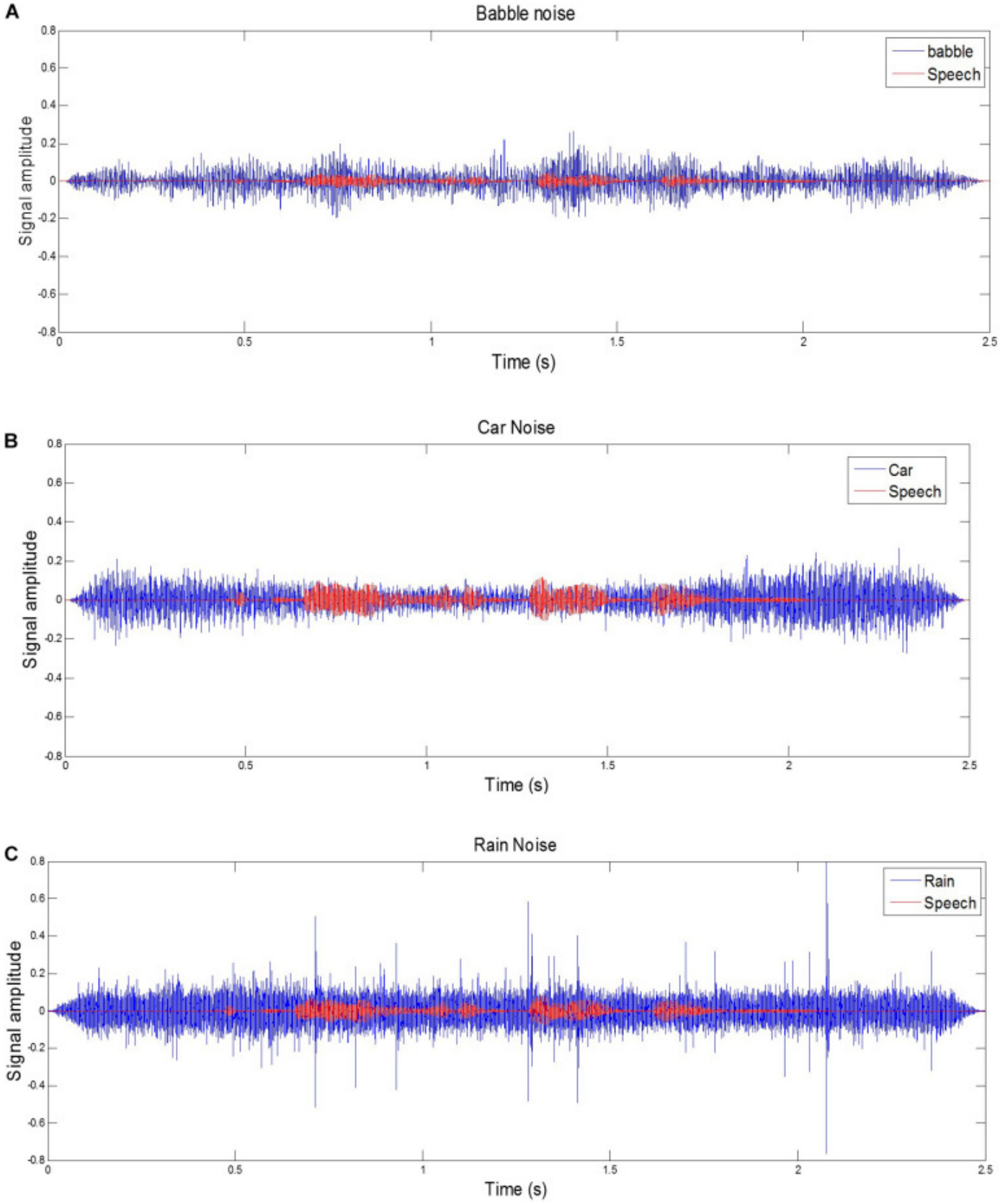
- **Accents and dialects:** Accents and dialects differ in pronunciation, vocabulary, and grammar, making it difficult for speech recognition applications to recognize speech accurately.
    - Assume a speech recognition model has been primarily trained on American English accents. If a speaker with a strong Scottish accent uses the system, they may encounter difficulties due to pronunciation differences. For example, the word "water" is pronounced differently in both accents. If the system is not familiar with this pronunciation, it may struggle to recognize the word "water."

Solution: Addressing these challenges is crucial to enhancing speech recognition applications' accuracy. To overcome pronunciation variations, it is essential to expand the training data to include samples from speakers with

diverse accents. This approach helps the system recognize and understand a broader range of speech patterns.

- **Background noise:** Background noise (e.g., traffic, cross-talk) makes distinguishing speech from background noise difficult for speech recognition applications (Figure 3).
  Solution: Pre-processing techniques can be used to reduce background noise in speech recognition, which can help improve the performance of speech recognition models in noisy environments.
  - For instance, you can use data augmentation techniques to reduce the impact of noise on audio data. Data augmentation helps train speech recognition models with noisy data to improve model accuracy in real-world environments.

Figure 3: Examples of a target sentence (“The clown had a funny face”) in the background noise of babble, car and rain.

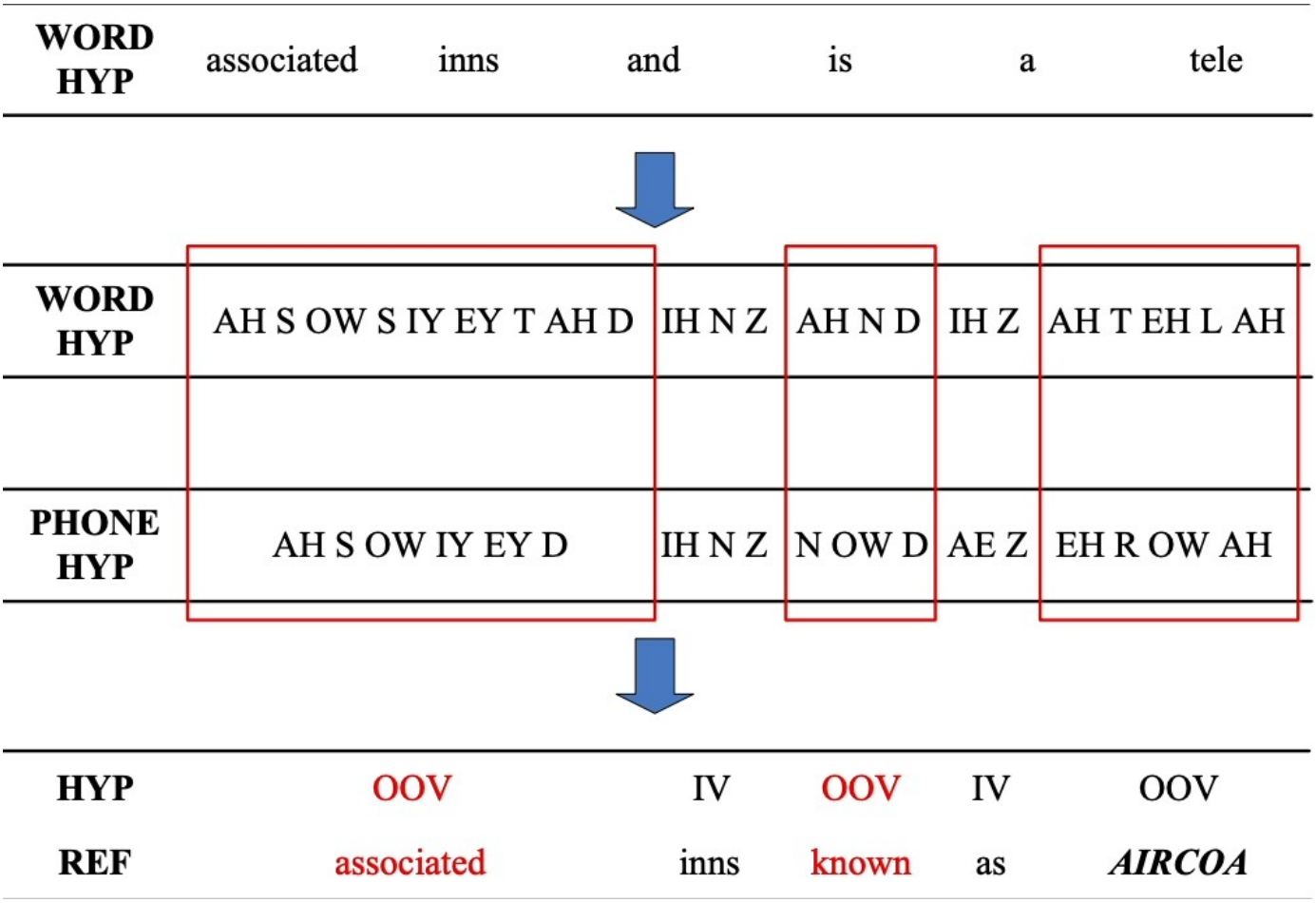


Source: PubMed Central[2]

Linguistic Challenges:

- **Out-of-vocabulary words:** Since the speech recognizers model has not been trained on OOV words, they may incorrectly recognize them as different or fail to transcribe them when encountering them.

Figure 4: An example of detecting OOV word



Source: Learning Out-of-vocabulary Words in Automatic Speech Recognition. [3]

Solution: Word Error Rate (WER) is a common metric that is used to measure the accuracy of a speech recognition or machine translation system. The word error rate can be computed as:

Figure 5: Demonstrating how to calculate word error rate (WER)

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- C is the number of correct words,
- N is the number of words in the reference (N=S+D+C)

Source: Wikipedia[4]

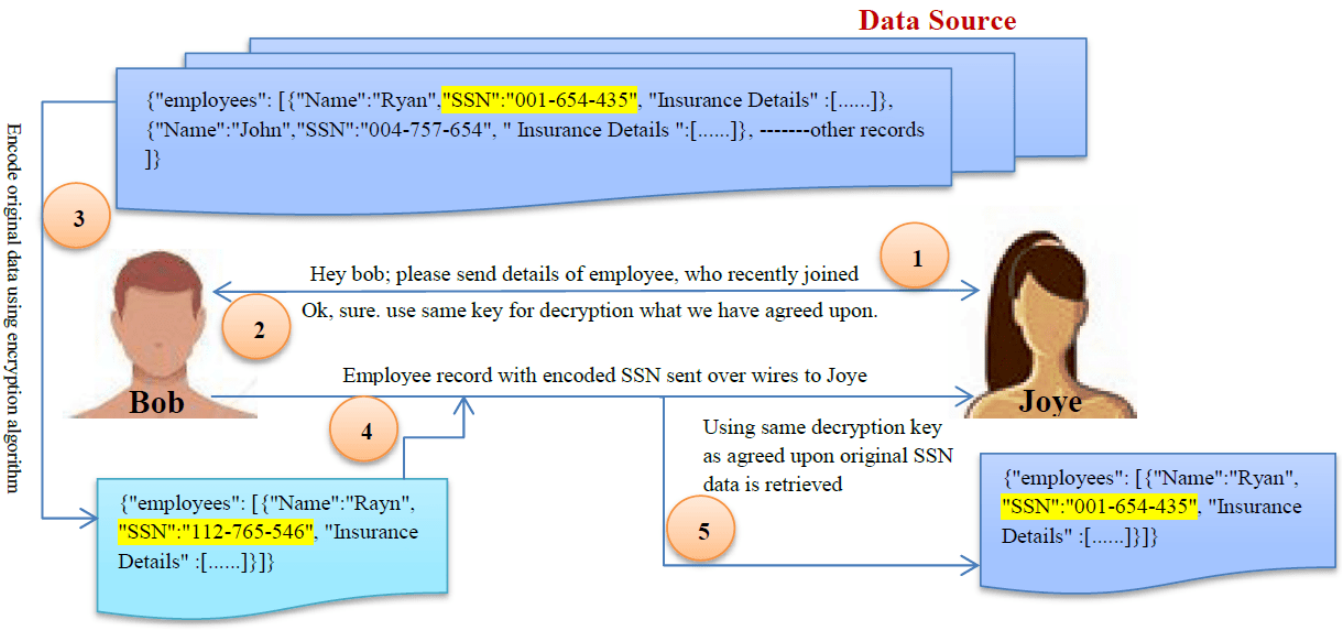
- **Homophones:** Homophones are words that are pronounced identically but have different meanings, such as “to,” “too,” and “two”.
  Solution: Semantic analysis allows speech recognition programs to select the appropriate homophone based on its intended meaning in a given context. Addressing homophones improves the ability of the speech recognition process to understand and transcribe spoken words accurately.

Technical/System Challenges:

- **Data privacy and security:** Speech recognition systems involve processing and storing sensitive and personal information, such as financial information. An unauthorized party could use the captured information, leading to privacy breaches.

Solution: You can encrypt sensitive and personal audio information transmitted between the user’s device and the speech recognition software. Another technique for addressing data privacy and security in speech recognition systems is data masking. Data masking algorithms mask and replace sensitive speech data with structurally identical but acoustically different data.

Figure 6: An example of how data masking works



Encryption and decryption cycle using symmetric encryption algorithm

Source: Informatica

- **Limited training data:** Limited training data directly impacts the performance of speech recognition software. With insufficient training data, the speech recognition model may struggle to generalize different accents or recognize less common words.

Solution: To improve the quality and quantity of training data, you can expand the existing dataset using data augmentation and synthetic data generation technologies.

13 speech recognition use cases and applications

In this section, we will explain how speech recognition revolutionizes the communication landscape across industries and changes the way businesses interact with machines.

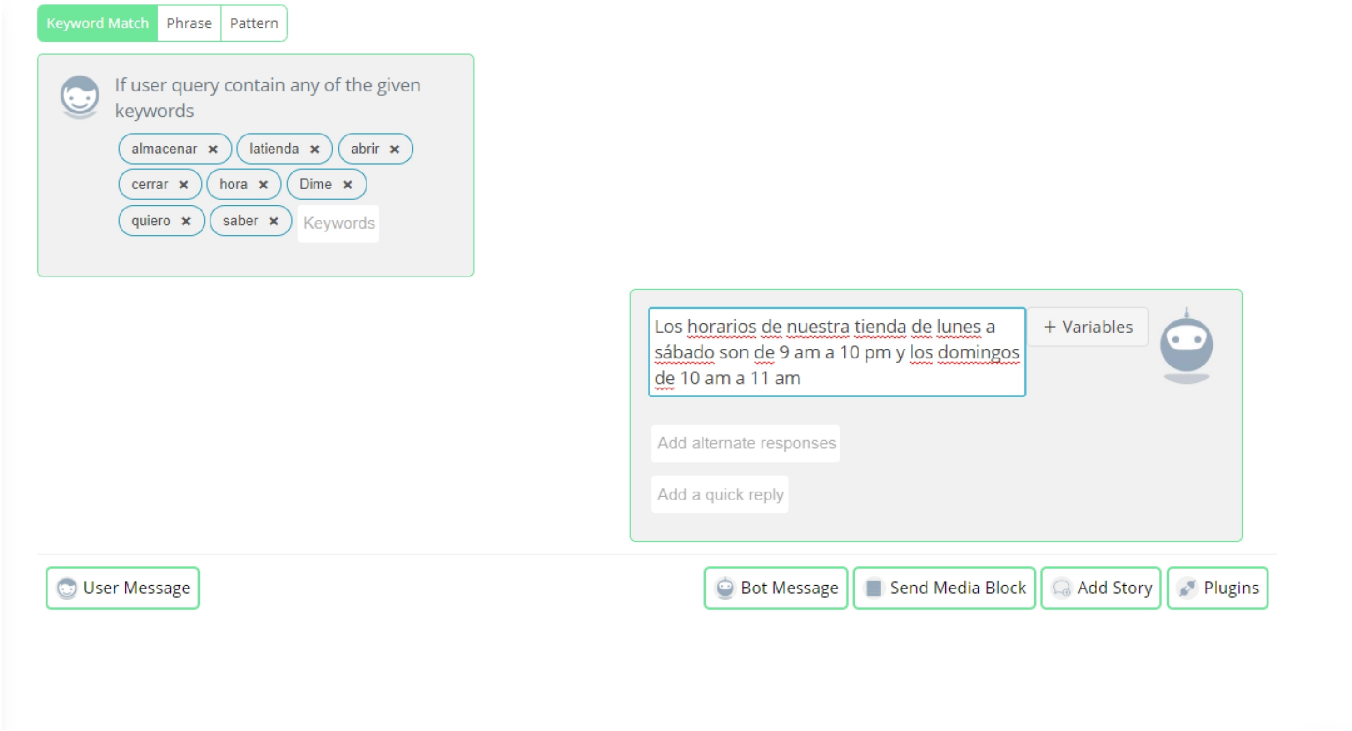
Customer Service and Support

1. **Interactive Voice Response (IVR) systems:** Interactive voice response (IVR) is a technology that automates the process of routing callers to the appropriate department. It understands customer queries and routes calls to the relevant departments. This reduces the call volume for contact centers and minimizes wait times. IVR systems address simple

customer questions without human intervention by employing pre-recorded messages or text-to-speech technology. Automatic Speech Recognition (ASR) allows IVR systems to comprehend and respond to customer inquiries and complaints in real time.

2. **Customer support automation and chatbots:** According to a survey, 78% of consumers interacted with a chatbot in 2022, but 80% of respondents said using chatbots increased their frustration level.
3. **Sentiment analysis and call monitoring:** Speech recognition technology converts spoken content from a call into text. After speech-to-text processing, natural language processing (NLP) techniques analyze the text and assign a sentiment score to the conversation, such as positive, negative, or neutral. By integrating speech recognition with sentiment analysis, organizations can address issues early on and gain valuable insights into customer preferences.
4. **Multilingual support:** Speech recognition software can be trained in various languages to recognize and transcribe the language spoken by a user accurately. By integrating speech recognition technology into chatbots and Interactive Voice Response (IVR) systems, organizations can overcome language barriers and reach a global audience (Figure 7). Multilingual chatbots and IVR automatically detect the language spoken by a user and switch to the appropriate language model.

Figure 7: Showing how a multilingual chatbot recognizes words in another lan guage



Source: Botsify

5. **Customer authentication with voice biometrics:** Voice biometrics use speech recognition technologies to analyze a speaker's voice and extract features such as accent and speed to verify their identity.

Sales and Marketing:

6. **Virtual sales assistants:** Virtual sales assistants are AI-powered chatbots that assist customers with purchasing and communicate with them through voice interactions. Speech recognition allows virtual sales assistants to understand the intent behind spoken language and tailor their responses based on customer preferences.
7. **Transcription services:** Speech recognition software records audio from sales calls and meetings and then converts the spoken words into written text using speech-to-text algorithms.

Automotive:

8. **Voice-activated controls:** Voice-activated controls allow users to interact with devices and applications using voice commands. Drivers can operate features like climate control, phone calls, or navigation systems.
9. **Voice-assisted navigation:** Voice-assisted navigation provides real-time voice-guided directions by utilizing the driver's voice input for the destination. Drivers can request real-time traffic updates or search for nearby points of interest using voice commands without physical controls.

Healthcare:

10. **Medical transcription:**Medical transcription, also known as MT, is the process of converting voice-recorded medical reports into a written text document. The following are the main steps in the medical transcription process:
    - Recording the physician's dictation
    - Transcribing the audio recording into written text using speech recognition technology
    - Editing the transcribed text for better accuracy and correcting errors as needed
    - Formatting the document in accordance with legal and medical requirements.

11. **Virtual medical assistants:** Virtual medical assistants (VMAs) use speech recognition, natural language processing, and machine learning algorithms to communicate with patients through voice or text. Speech recognition software allows VMAs to respond to voice commands, retrieve information from electronic health records (EHRs) and automate the medical transcription process.
12. **Electronic Health Records (EHR) integration:** Healthcare professionals can use voice commands to navigate the EHR system, access patient data, and enter data into specific fields.

Technology:

13. **Virtual agents:** Virtual agents utilize natural language processing (NLP) and speech recognition technologies to understand spoken language and convert it into text. Speech recognition enables virtual agents to process spoken language in real-time and respond promptly and accurately to user voice commands.

Further reading

- Top 5 Speech Recognition Data Collection Methods in 2023
- Top 11 Speech Recognition Applications in 2023

External Links

1. Databricks

2. PubMed Central

3. Qin, L. (2013). Learning Out-of-vocabulary Words in Automatic Speech Recognition. Carnegie Mellon University.

4. Wikipedia

Cem Dilmegani
Principal Analyst

Cem has been the principal analyst at AIMultiple since 2017. AIMultiple informs hundreds of thousands of businesses (as per similarWeb) including 60% of Fortune 500 every month.

Cem's work has been cited by leading global publications including Business Insider, Forbes, Washington Post, global firms like Deloitte, HPE, NGOs like World Economic Forum and supranational organizations like European Commission. You can see more reputable companies and media that referenced AIMultiple.

Throughout his career, Cem served as a tech consultant, tech buyer and tech entrepreneur. He advised businesses on their enterprise software, automation, cloud, AI / ML and other technology related decisions at McKinsey & Company and Altman Solon for more than a decade. He also published a McKinsey report on digitalization.

He led technology strategy and procurement of a telco while reporting to

the CEO. He has also led commercial growth of deep tech company Hypatos that reached a 7 digit annual recurring revenue and a 9 digit valuation from 0 within 2 years. Cem's work in Hypatos was covered by leading technology publications like TechCrunch and Business Insider.

Cem regularly speaks at international technology conferences. He graduated from Bogazici University as a computer engineer and holds an MBA from Columbia Business School.
Stay up-to-date on B2B tech, accelerate your enterprise

Follow on

Next to Read

Top 5 Speech Recognition Data Collection Methods in 2024

Jan 2 | 5 min read

Top 4 Speech Recognition Challenges & Solutions in 2024

Jan 3 | 4 min read

Top 11 Voice Recognition Applications in 2024

Jan 3 | 4 min read

Comments

0 Comments

Your email address will not be published. All fields are required.

Name Surname

Email Address

Write your comment..

Post Comment

AIMultiple

AIMultiple

Solutions

For Tech Users

Vendors

Investors

Enterprises use AIMultiple to identify new software and services, their use cases, benefits, best practices and case studies. AIMultiple shares data-driven insights on how solutions in AI / generative AI / machine learning / data science, cloud / cloud GPUs, cybersecurity / application security / network security / microsegmentation, data collection / web data / survey software, IoT, process mining, RPA / AP automation / workload automation / MFT can transform businesses.

Data-driven, Transparent, Practical New Tech Industry Analysis