Conversational AI

Steven Au sttau@ucsc.edu

What Is Conversational AI?

Conversational AI is the use of machine learning to develop speech-based apps that allow humans to interact naturally with devices, machines, and computers using audio. You use conversational AI when getting weather updates from your virtual assistant, when asking your navigation system for directions, or when communicating with a chatbot online. You speak in your normal voice and the device understands, finds the best answer, and replies with speech that sounds natural.

The Two Components of the Conversational AI Pipeline

1. Speech AI
   a. Automatic Speech Recognition (ASR) or Speech-to-Text (STT)
   b. Text-to-Speech (TTS) with voice synthesis
2. Natural Language Processing (NLP) or Natural Language Understanding (NLU)

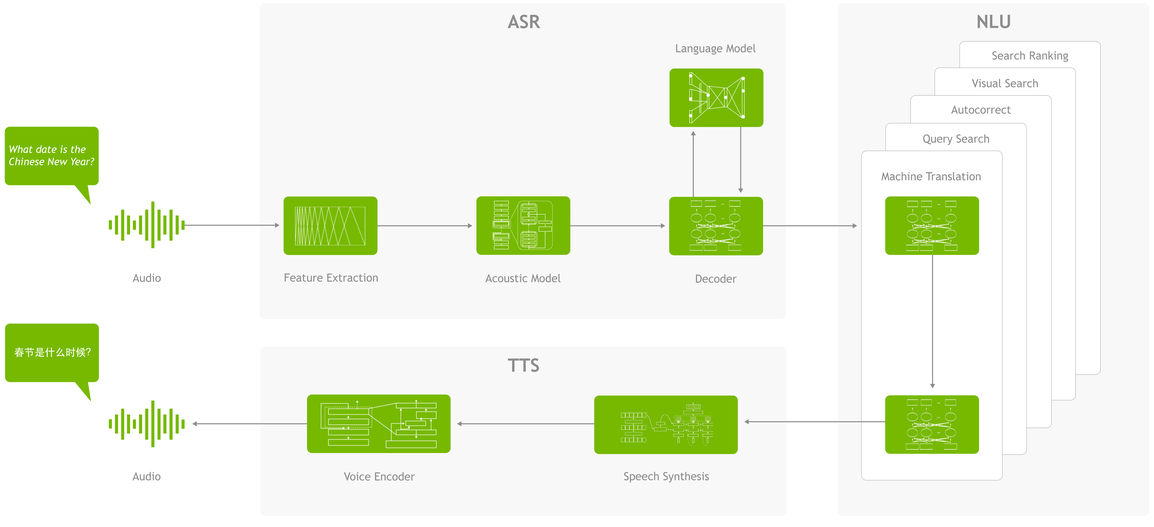


Figure 1: Overview of a conversational AI pipeline

1. The audio waveform is converted to text during the automatic speech recognition (ASR) stage.
2. The question is then interpreted, and the device generates a smart response during the natural language processing (NLP) stage.
3. Finally, the text is converted into speech signals to generate audio for the user during the text-to-speech (TTS) stage.

Several deep learning models are connected into a pipeline to build a conversational AI application.

NVIDIA AI Technolgies

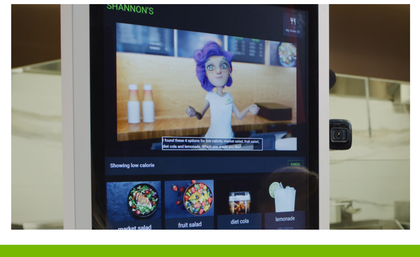
Riva - Speech AI

NeMo Megatron - Natural Language Processing

Tokki - AI-powered customer service agents

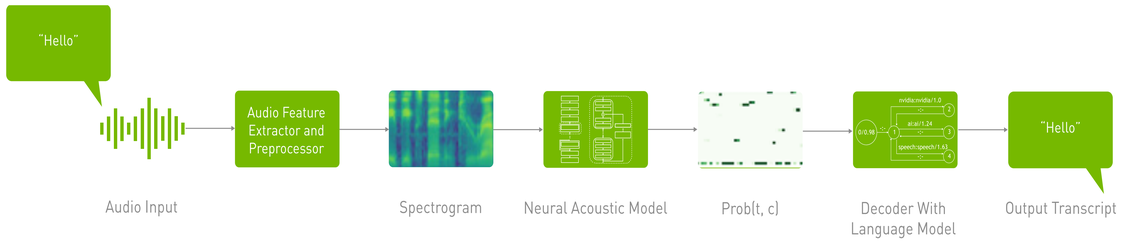
Chatbots are commonly used in retail applications to accurately understand customer queries and generate responses and recommendations

https://developer.nvidia.com/conversational-ai



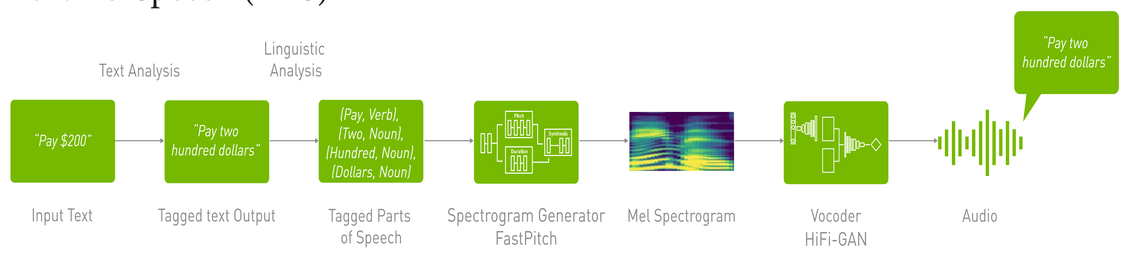
Speech AI

Automatic Speech Recognition (ASR)



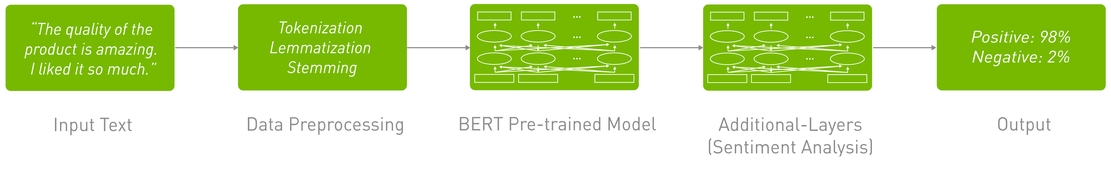
- Extract useful audio features from the input audio and ignore noise and other irrelevant information.
- Mel-frequency cepstral coefficient (MFCC) techniques capture audio spectral features in a spectrogram.
- The decoder and language model convert these characters into a sequence of words based on context.

Text-To-Speech (TTS)



- A synthesis network generates a spectrogram from text, and a vocoder network generates a waveform
- The output from text analysis is passed into linguistic analysis for refining pronunciations, calculating the duration of words, deciphering the prosodic structure of utterance, and understanding grammatical information.
- Output from linguistic analysis is then fed to a speech synthesis neural network model, which converts the text to mel spectrograms and then to a neural vocoder model to generate the natural-sounding speech.

Natural Language Understanding (NLU)



- Text is converted into an encoded vector using techniques such as Word2Vec, TF-IDF vectorization, and word embedding.
- These vectors are passed to a deep learning model, such as a recurrent neural network (RNN), long short-term memory (LSTM), and Transformer to understand context.