shagrawal Microsoft

Sep 03 2024 0

...

Next-Gen Voice Bots: Human-Like Interaction with Azure Speech

Co-authors: Kenichiro Nakamura

Introduction: Navigating the Evolution of Next-Gen Voice Bots

The Importance of Voice Conversations

AI-powered voice conversations are becoming a critical priority in the increasingly competitive landscape of customer interaction. With the rise of digital players, organizations recognize the power of voice bots as a natural and intuitive mode of communication that can deliver human-like experiences, deeply engaging users and setting them apart from competitors. The growing demand for high-quality voice interactions is fueled by the need for seamless customer service, personalized assistance, and instant information access. Moreover, as companies strive to retain and expand revenue, reaching a more diverse customer base across language barriers has become essential, making multilingual and context-aware voice solutions a key differentiator in today's market.

Key Challenges in Creating Effective Voice Bot Solutions

Despite the potential, creating voice bot solutions that truly resonate with users is fraught with challenges. Very few organizations have successfully addressed the key hurdles that hinder the development of state-of-the-art voice bots:

- **Latency:** Ensuring that voice interactions happen in real-time, without noticeable delays, is critical for maintaining natural conversations. High latency can disrupt the flow of dialogue, leading to user frustration and diminished engagement.
- **Accuracy:** Accurate speech recognition is essential, especially in noisy environments or with users who have diverse accents and dialects. Misinterpretation of spoken words can lead to incorrect responses and a breakdown in communication.
- **Cost Efficiency:** Organizations face the challenge of creating an architecture that balances advanced functionality with cost-effective operations, and thus struggle to see the ROI on their investments.
- **Personalized, Human-Like Conversations**: Users expect voice bots to understand context, display empathy, and provide responses that feel personal and relatable. Achieving this level of interaction requires careful selection of the right LLM from the many options available today, along with implementing custom voice capabilities to enhance the conversational experience.

Empowering Human-Like Interactions Through Next-Gen Voice Bots

In following sections, we'll explore how addressing these core challenges with Azure AI capabilities can empower businesses to deliver next-gen voice experiences that exceed customer expectations. Here is a quick demo leveraging some of the capabilities in Azure AI Stack to showcase a voice bot engaging is promotional sales conversations:

Enhancing Accuracy

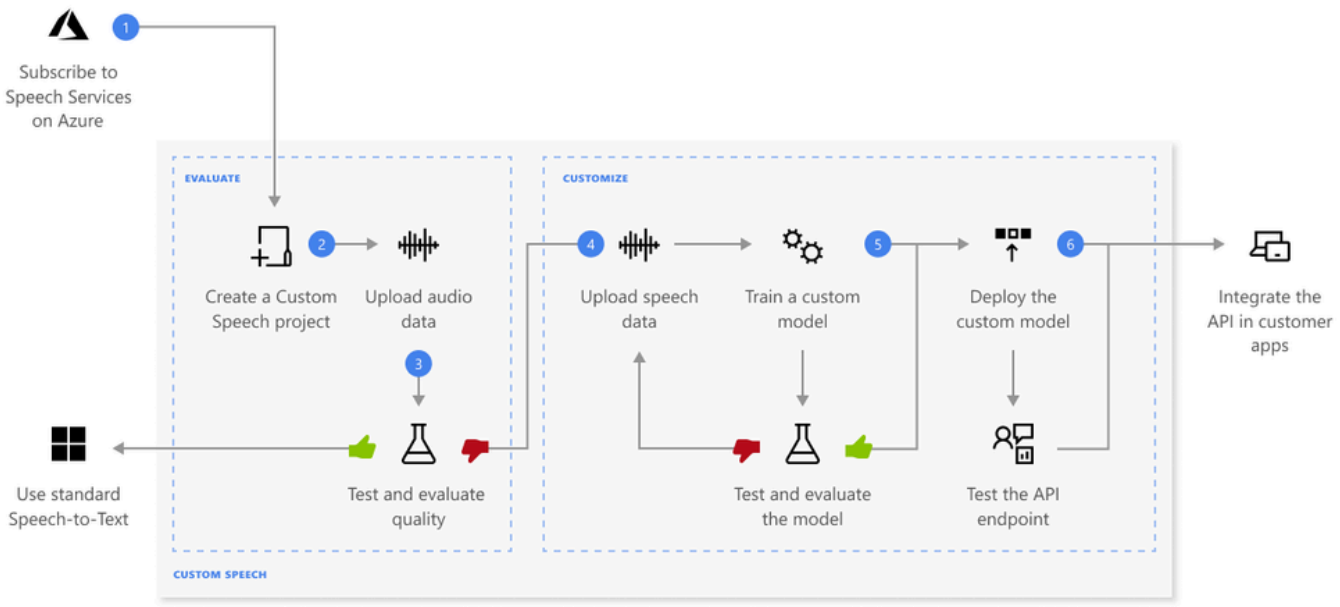
Custom Speech Models for Diverse Scenarios

Azure Custom Speech service enables businesses to fine-tune Automatic Speech Recognition (ASR) for specific needs by leveraging domain-specific vocabulary, pronunciation guides, and tailored acoustic environments. These customizations enhance speech recognition accuracy and improve user experiences across various use cases.

Key Capabilities of Custom Speech Models

1. **Handling Noise and Acoustic Variations**: Custom Speech models can be trained to maintain accuracy in noisy environments and varying acoustic conditions, such as busy streets, public spaces, or drive-throughs. By using data augmentation techniques, like mixing clean audio with background noise, models can become robust against diverse soundscapes.
2. **Domain-Specific Vocabulary**: Improve recognition of industry-specific jargon and technical terms. Custom Speech can accurately handle specialized language in fields like healthcare, legal, and finance, ensuring that conversations involving complex terms are transcribed correctly. Example: Recognizing specialized scientific terms or product names accurately during technical presentations or customer support calls.

3. **Custom Pronunciation**: Tailor models to recognize non-standard pronunciations and unique terms, such as brand names or regional dialects, ensuring accurate transcription of spoken words.
4. **Accent and Language Support**: Adapt models to recognize various accents and dialects, enhancing global accessibility and user engagement.
5. **Enhanced Output Formatting**: Define specific text formatting rules, such as number normalization and profanity filtering, to meet industry standards for clarity and appropriateness.



Use Cases

- **Education:** Accurate live captioning during academic lectures.
- **Healthcare:** Reliable transcription of medical consultations.
- **Customer Support:** Improved accuracy for call centers handling diverse accents.
- **Media:** Accurate reporting of names and places in live broadcasts.

Call to Action: Leverage Azure Custom Speech to enhance your voice-enabled applications. Address challenges like noise, complex terminology, and accents to provide a seamless, engaging user experience.

Ref Links:

Custom speech overview - Speech service - Azure AI services | Microsoft Learn

Speech Studio - Custom Speech - Overview (microsoft.com)

Sample Training Data for Custom Model Finetuning

https://docs.nvidia.com/deeplearning/riva/user-guide/docs/tutorials/asr-noise-augmentation-offline.h...

Personal Voice Creation

Customizing AI Voices

Azure AI text to speech enables developers to convert text into human like synthesized speech. Neural TTS is a text to speech system that uses deep neural networks to make the voices of computers nearly indistinguishable from recordings of people. It provides human-like natural prosody and clear articulation of words, which significantly reduces listening fatigue during interaction with AI systems. With Azure personal voice feature, users can create customized AI voices that replicate their own or specific personas. By providing a brief speech sample, you can generate a unique voice model capable of synthesizing speech in over 90 languages across more than 100 locales. This functionality is particularly beneficial for applications like personalized virtual assistants, enhancing user engagement and interaction by utilizing voices that are familiar and relatable to the audience. Once created, personal voice can used in application using ssml:

```
if blnPersonalVoice:
        speaker_profile_id = "e04805d2-b81c-48ed-ac6b-1fa099edf0f3"

        ssml = "<speak version='1.0' xml:lang='hi-IN' xmlns='http://www.w3.org/2001/10/syn
       "xmlns:mstts='http://www.w3.org/2001/mstts'>" \
       "<voice name='DragonLatestNeural'>" \
       "<mstts:ttsembedding speakerProfileId='%s'/>" \
       "<mstts:express-as style='cheerful' styledegree='5'>" \
       "<lang xml:lang='%s'> %s </lang>" \
       "</mstts:express-as>" \
       "</voice></speak> " % (speaker_profile_id, locale, text)
        result_future = synthesizer.speak_ssml_async(ssml)
    else:
        result_future = synthesizer.speak_text_async(text)


    result = await loop.run_in_executor(None, result_future.get)
```

Call To Action: Explore how to implement personalized voice features in your applications to enhance user experience and engagement!

Ref Links:

https://learn.microsoft.com/en-us/azure/ai-services/speech-service/personal-voice-overview

cognitive-services-speech-sdk/samples/custom-voice at master · Azure-Samples/cognitive-services-spee...

Voice and sound with Speech Synthesis Markup Language (SSML) - Speech service - Azure AI services | ...

Achieving Low Latency with Real-Time Audio Synthesis

To deliver seamless, low-latency voice interactions, leveraging real-time audio synthesis with Azure Speech SDK and OpenAI's streaming capabilities is essential. By processing responses in small chunks and synthesizing audio as soon as each chunk is ready, you can provide a fluid, conversational experience.

Stream Responses from Azure OpenAI

Start by streaming text responses from OpenAI in real time:

- **Stream Responses**: Use OpenAI's streaming capability to receive partial text responses as they are generated.
- **Buffer and Process**: Accumulate text until a complete thought (indicated by punctuation) is detected, then initiate synthesis.

```
completion = client.chat.completions.create(model=open_ai_deployment_name, messages=messag

async def process_stream():
    text_buffer = ""
    for event in completion:
        if choice := event.choices[0].delta.content:
            text_buffer += choice
            if any(p in text_buffer for p in ",;.!?"):
                await text_to_speech_streaming(text_buffer.strip())
                text_buffer = ""  # Clear buffer
```

Set Up Audio Output with Push Model

Use the push model to stream audio data as soon as it is synthesized:

```
# Custom class to handle pushed audio data
class CustomPushAudioStream(PushAudioOutputStreamCallback):
    def write(self, audio_buffer: memoryview) -> int:
        # Handle the received audio data (e.g., play it, save it)
        print(f"Received audio buffer of size: {len(audio_buffer)}")
        return len(audio_buffer)

# Create a global SpeechSynthesizer with custom push stream
push_stream = CustomPushAudioStream()
audio_config = AudioConfig(stream=push_stream)
synthesizer = SpeechSynthesizer(speech_config=speech_config, audio_config=audio_config)

# Function to perform text-to-speech synthesis
async def text_to_speech_streaming(text):
    result = synthesizer.speak_text_async(text).get()
    if result.reason == ResultReason.SynthesizingAudioCompleted:
        print(f"Synthesis complete for: {text}")
    elif result.reason == ResultReason.Canceled:
        print("Synthesis canceled.")
```

Call To Action: By first streaming responses from OpenAI and then immediately pushing audio output to playback, you can achieve low latency and high responsiveness in voice interactions. This push-based streaming approach is ideal for real-time, dynamic conversations, ensuring a natural and engaging user experience.

Ref Links:

Make your voice chatbots more engaging with new text to speech features (microsoft.com)

How to lower speech synthesis latency using Speech SDK - Azure AI services | Microsoft Learn

User Experience Boost

Smart Prompts with OpenAI Integration

The integration of OpenAI with Azure AI Speech enhances user experience through smart prompts, making interactions more engaging and personalized. Leveraging natural language processing capabilities, these systems understand context and generate relevant responses in real-time, enabling seamless conversations in customer support or virtual assistant scenarios. Additionally, by instructing OpenAI to include punctuation, voice bots can leverage streaming capabilities to generate audio responses with appropriate pauses and intonation. This not only makes interactions more natural but also reduces latency by playing back audio incrementally as it's being synthesized, enhancing the overall user experience.

```
**Conversation Protocol**
        1. You converse with customer in simple, short , sentences.
        2. You use punctuations frequently - ,;.!?
        3. You generate text so that in the begining you have a small phrase ending in pun
```

Call To Action: Discover how integrating smart prompts into your applications can elevate customer interactions and streamline communication processes!

Achieving Low Latency with Real-Time Speech-to-Text Streaming

Real-time speech-to-text (STT) streaming using Azure Speech SDK with `PushAudioInputStream` enables immediate transcription of speech, providing a responsive and natural user experience. This approach is ideal for scenarios requiring quick feedback, such as customer support, live transcription, and interactive voice systems.

Main Benefit

Immediate Feedback: Using `PushAudioInputStream` for real-time STT ensures speech is transcribed as soon as it's spoken, maintaining the flow of conversation and enhancing the overall user experience.

```

speech_config = speechsdk.SpeechConfig(subscription=speech_key, region=speech_region)

# Create a push audio input stream and audio configuration
stream = speechsdk.audio.PushAudioInputStream()
audio_config = speechsdk.audio.AudioConfig(stream=stream)

# Create the SpeechRecognizer with push stream input
speech_recognizer = speechsdk.SpeechRecognizer(language=lang, speech_config=speech_config,

# Global list to store recognized text
text = []

# Callback function to handle recognized speech
def handle_recognized(evt):
    if evt.result.reason == speechsdk.ResultReason.RecognizedSpeech:
        text.append(evt.result.text)
        print(f"Recognized: {evt.result.text}")

# Connect the callback function to the recognized event
speech_recognizer.recognized.connect(handle_recognized)

# Start continuous recognition
speech_recognizer.start_continuous_recognition()
```

Ref Links:
Speech SDK audio input stream concepts - Azure AI services | Microsoft Learn

Real-Time Interruption handling with Streaming Architecture

In conversational AI, handling interruptions gracefully is essential for creating a natural dialogue flow. With a streaming architecture, voice bots can detect and respond to user interruptions in real-time. By continuously monitoring for human speech while streaming bot responses, the system can immediately stop playback as soon as it detects a user speaking. This ensures that the bot does not continue talking over the user, making interactions more natural and less frustrating. Utilizing Azure Speech SDK's real-time capabilities allows developers to build bots that not only stop the TTS stream on user input but also accurately manage conversation context and seamlessly switch back to listening mode, enhancing overall user experience.
Call To Action: how implementing real-time interruption handling in your voice bot can create more natural and responsive interactions, leading to higher user satisfaction!

Speaker Identification with Real-Time Diarization

Real-time diarization is a powerful feature that differentiates speakers in audio streams, enabling systems to recognize and transcribe speech segments attributed to specific speakers. This capability is particularly beneficial in scenarios like meetings or multi-participant discussions, where knowing who said what can enhance clarity and understanding. By employing single-channel audio streaming, the technology can accurately identify distinct voices and associate them with the respective dialogue, providing a structured transcription output that includes speaker labels.
Call To Action: Explore how integrating real-time diarization can elevate your call center operations by improving call analytics and enhancing customer interactions!
Ref Links:
https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/announcing-general-availability-of-...

Multilingual Capabilities

Automatic Language Detection and Translation

Azure automatic language detection and translation features significantly enhance user interactions by enabling real-time translations without the need for users to specify input languages. This capability allows applications to seamlessly identify spoken languages, facilitating communication in multilingual scenarios. The Speech Translation API can handle multiple languages in a single session, automatically switching between them as needed while providing accurate translations in text or audio form. Further Azure AI text to speech offers more than 400 voices and more than 140 languages and locales. A single pre-built realistic neural voice with multilingual support makes it easy to read content in a broad range of languages in the same voice.

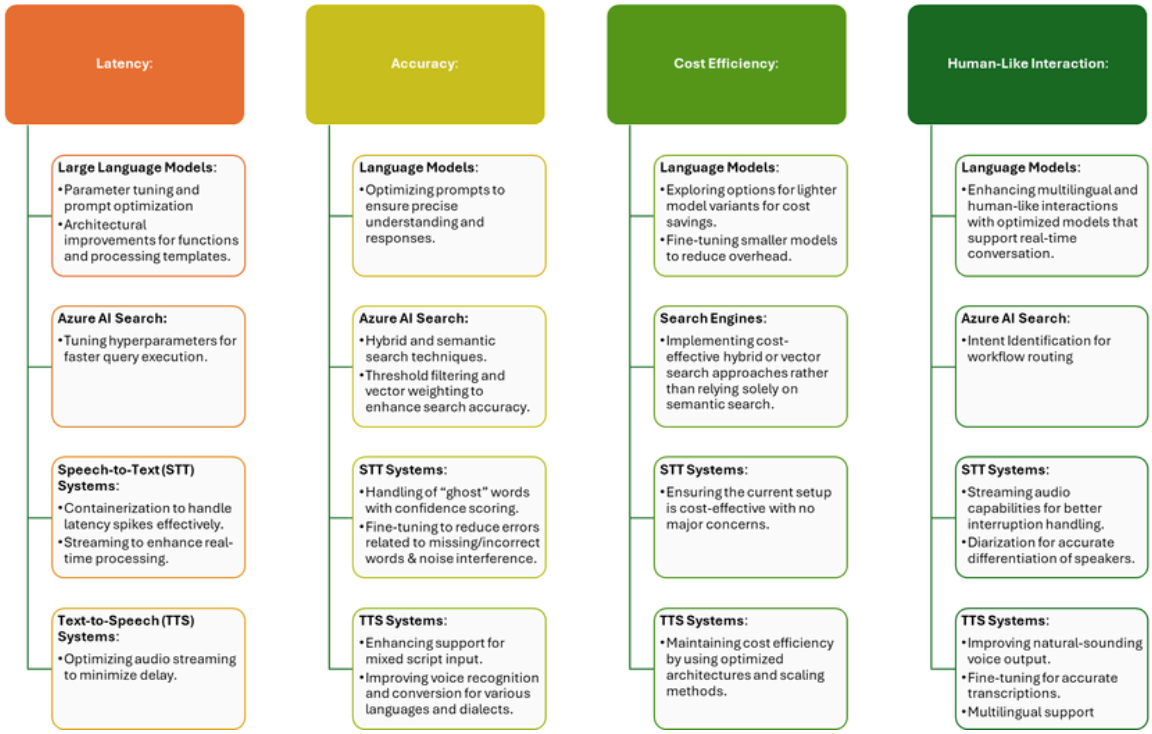
Call To Action: Discover how integrating automatic language detection and translation can elevate your customer interactions across diverse markets!

Ref Links: Announcing-video-translation-and-speech-translation-api

Conclusion

The Path to Success Powered by Innovations in Azure AI

Innovations in Azure AI Speech, Azure AI Speech and Azure Open AI are paving the way for continuous success stories in the realm of voice bots.



Azure cutting-edge technologies provide comprehensive solutions to key challenges in voice bot development. With low latency, high accuracy, cost-effective scaling, and human-like interactions, Azure empowers businesses to deliver responsive and engaging voice experiences that meet and exceed customer expectations. By leveraging these capabilities, organizations can enhance their communication strategies and drive meaningful user engagement.

References:

Make your voice chatbots more engaging with new text to speech features (microsoft.com)

Announcing-new-multi-modal-capabilities-with-azure-ai-speech

Guidebook-to-reduce-latency-for-azure-speeche-speech-to-text-stt-and/ba-p/4208289

 2 Likes