NYP NANYANG
THE INNOVATIVE POLYTECHNIC

AI Voice Agent Application

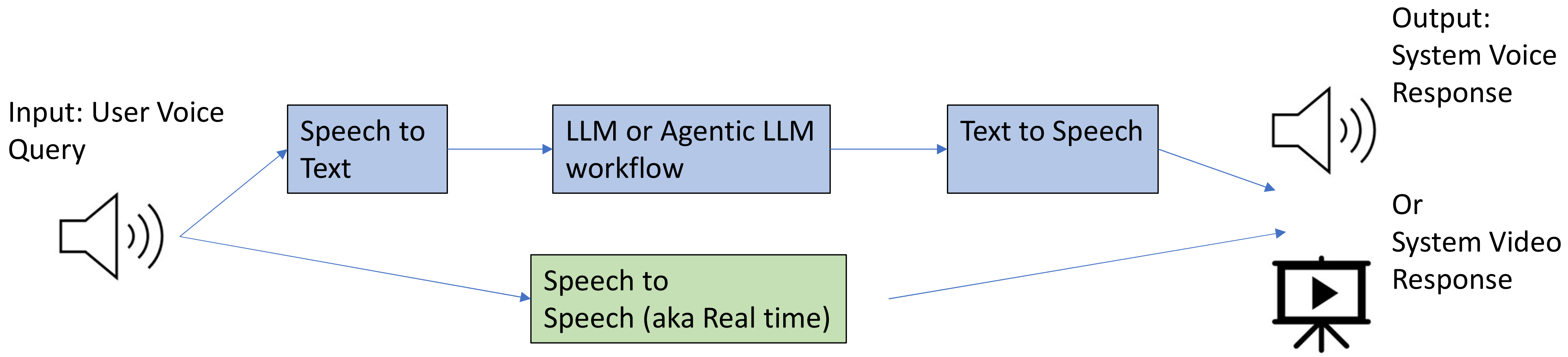
Learning Outcomes

Upon completion of this session, the learners should be able to:

- Understand the overview of AI Voice Agent Stack Components
- Apply the Stack Components to develop the AI Voice Agent application

What is an AI Voice Agent?

- Definition: Voice Agents combine speech and reasoning abilities of douncation models to deliver real-time, human-like voice interactions.
- Use Cases
  - Improve learning: Guide personalized skill development, conduct interviews
  - Handle customer service voice calls(Restaurant booking, sales, insurance)
  - Improve accessibility in medical and talk therapy application



Input: User Voice Query
Speech to Text
LLM or Agentic LLM workflow
Text to Speech
Speech to Speech (aka Real time)
Output: System Voice Response
Or System Video Response

Voice Agent Stack Components

- Automatic Speech Recognition(ASR)., also know as Speech-to- Text(STT): the task of transcribing a given audio signal to text.
- Audio -> Text

- LLMs and LLM agent: Generate a response to the transcribed query
- Text -> Text or multimodal response(e.g. text or images)

- Text-to-Speech(TTS), also know as Speech Synthesis: the task of generating natural and intelligible speech from text
- Text-> Audio

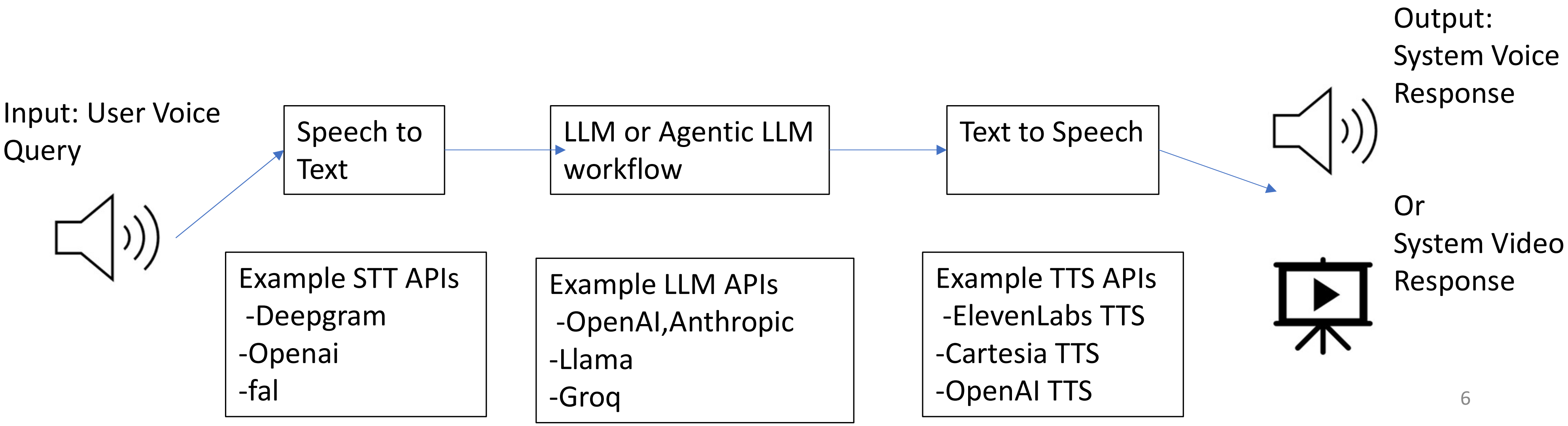
AI Voice Agent Stack Components

Other Consideration beyond AI Voice Stack Components

Voice Activity Detection (VAD)
Detecting presence/ absence of human speech in audio

End of Turn/Utterance detection (EOU)
Detecting whether a speaker has finished their turn in the conservation

Examples Providers: TTS/LLM/STT Pipeline



Input: User Voice Query

Speech to Text

LLM or Agentic LLM workflow

Text to Speech

Output: System Voice Response

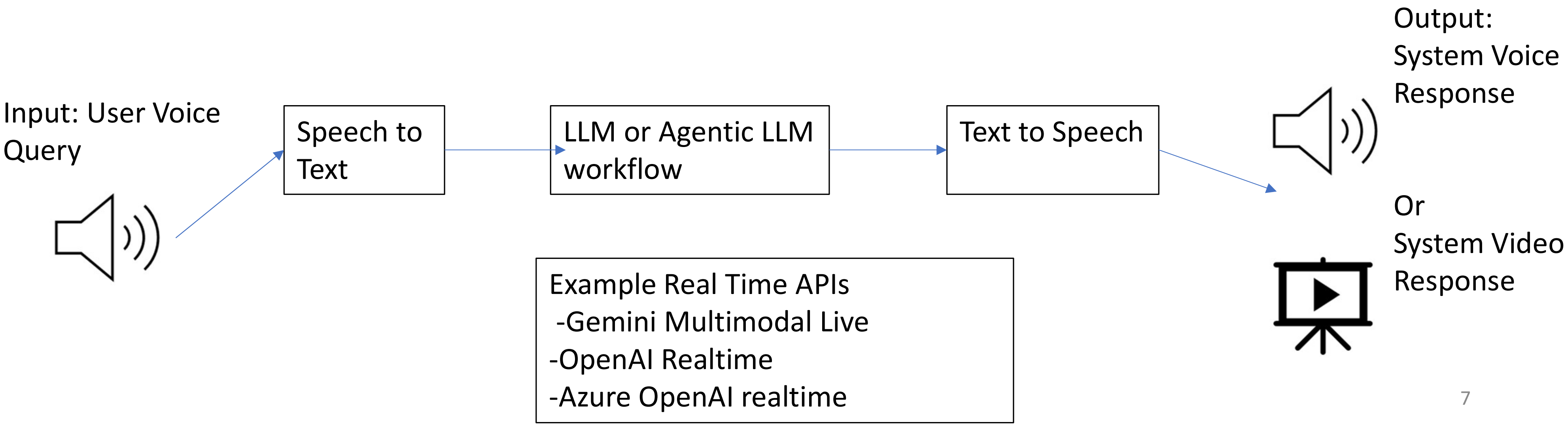
Or System Video Response

Example STT APIs
-Deepgram
-Openai
-fal

Example LLM APIs
-OpenAI,Anthropic
-Llama
-Groq

Example TTS APIs
-ElevenLabs TTS
-Cartesia TTS
-OpenAI TTS

Examples Providers: TTS/LLM/STT Pipeline(real time)



Input: User Voice Query

Speech to Text

LLM or Agentic LLM workflow

Text to Speech

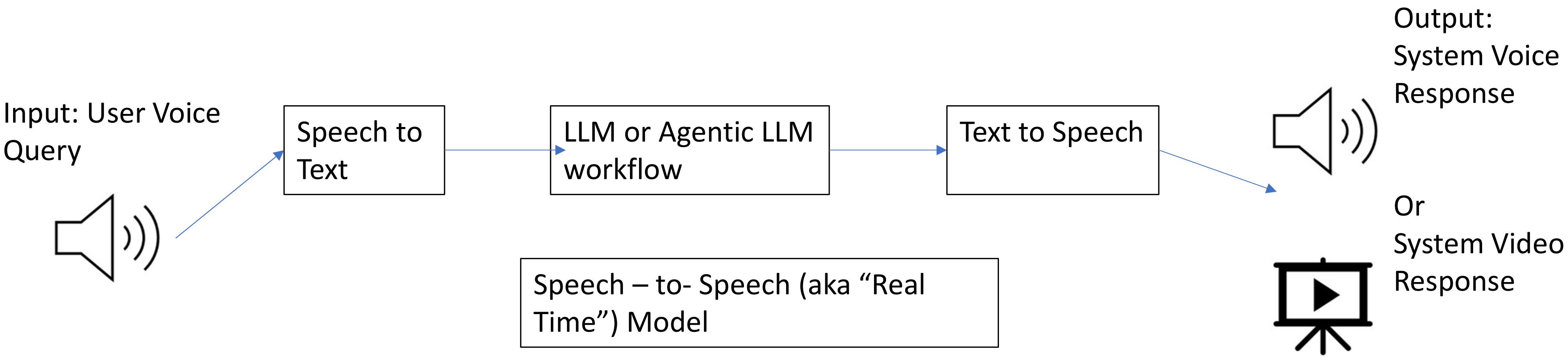
Output: System Voice Response

Or System Video Response

Example Real Time APIs
-Gemini Multimodal Live
-OpenAI Realtime
-Azure OpenAI realtime

Examples Providers: TTS/LLM/STT Pipeline(real time)

- Live Conversation requires low latency audio streaming



Input: User Voice Query

Speech to Text

LLM or Agentic LLM workflow

Text to Speech

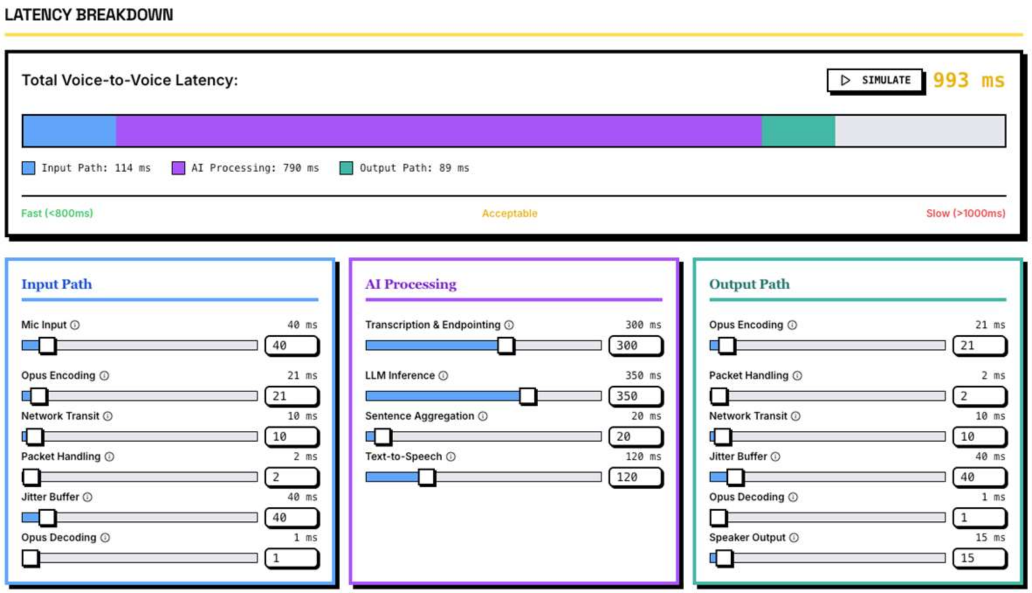
Output: System Voice Response

Or System Video Response

Speech – to- Speech (aka “Real Time”) Model

Understanding Voice-to-Voice Latency

- Latency is perhaps the most critical factor in voice agent performance. In human conversation, responses typically arrive within 500ms, and responses beyond 1000ms feel unnaturally delayed. For voice AI agents to feel natural, they must achieve similar responsiveness. This guide explores optimization strategies across all layers of the voice AI stack.



LATENCY BREAKDOWN

Total Voice-to-Voice Latency: 993 ms

🟦 Input Path: 114 ms 🟪 AI Processing: 790 ms 🟩 Output Path: 89 ms

Fast (<800ms) Acceptable Slow (>1000ms)

Input Path

Mic Input 40 ms — 40
Opus Encoding 21 ms — 21
Network Transit 10 ms — 10
Packet Handling 2 ms — 2
Jitter Buffer 40 ms — 40
Opus Decoding 1 ms — 1

AI Processing

Transcription & Endpointing 300 ms — 300
LLM Inference 350 ms — 350
Sentence Aggregation 20 ms — 20
Text-to-Speech 120 ms — 120

Output Path

Opus Encoding 21 ms — 21
Packet Handling 2 ms — 2
Network Transit 10 ms — 10
Jitter Buffer 40 ms — 40
Opus Decoding 1 ms — 1
Speaker Output 15 ms — 15

Latency Lower Bounds

- Latency of Human Interactions
- Humans expect a response on avg within 230ms(std dev=520ms) from the end of their interlocutor’s turn
- Note: estimates for English, other languages maybe be up or down

- Latency of Voice Agent interactions(Note: all streaming APIs for STT/LLM/TTS)

Task	Latency
VAD(LiveKit’s)	20ms
EOU(LiveKit’s)	100ms
ASR/STT	100-500ms
LLM (time to first token)	200-550ms
TTS (time to first byte)	100-450ms

Interlocutor: a person who takes part in a dialogue or conversation.

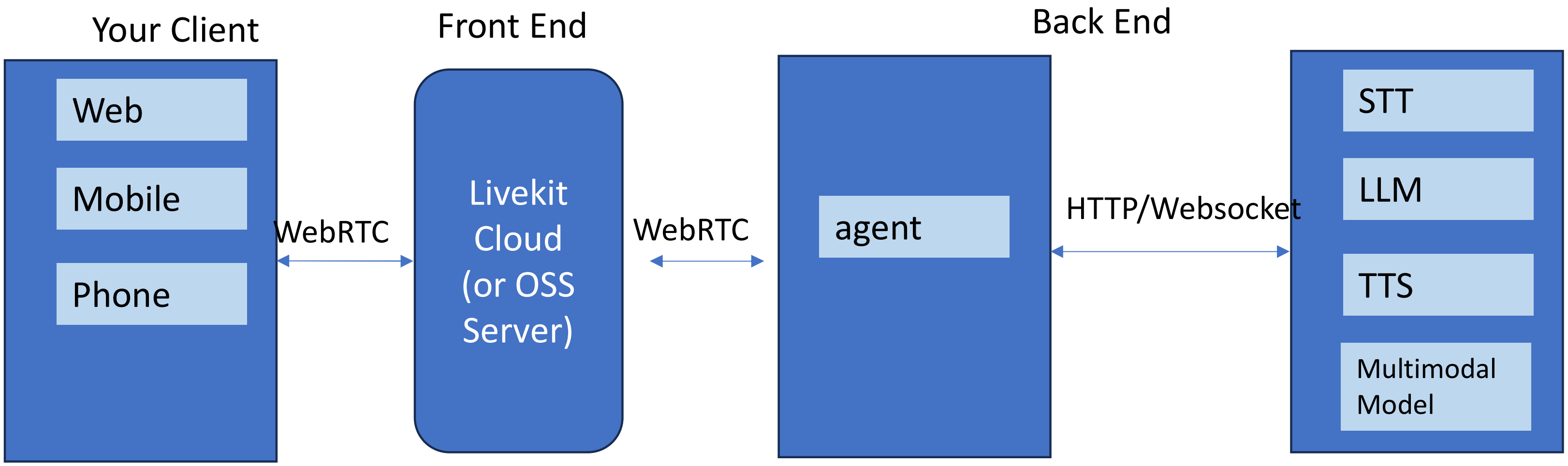
Evaluating and Optimizing Voice Agents

- Unique Challenges
  - Speech and Text artifacts due to ASR/TTS and VAD/EOU
  - Multilingual ASR performance typically lags behind English ASR.
- Latency Optimization
  - Non trivial to estimate in practice(client vs server-side measurement)
    - LiveKit/VAPI provides a low latency network
  - IN Voice Agents using the STT+LLM+TTS architecture, the LLM often the primary source of latency
    - If self hosting, use smaller/quantized model
    - If API access, consider API limits, provide recipes, etc
    - Shorten the replay(e.g. through LLM prompting) or construct the reply in segments(e.g. interstitial or a short acknowledgement before the fill reply is given)

Official (Open)

Approach: Real Time Peer-to-Peer Communication

- Web Real-Time Communication(WebRTC) is a free open source project providing web browser and mobile applications with real-time communication vis APIs.
- WebSocket network communication protocol to establish a client-server 'handshake'.
- Core elements: asynchronous processing and careful management of I/O stream & streaming APIs(ie. STT/TTS/LLM)



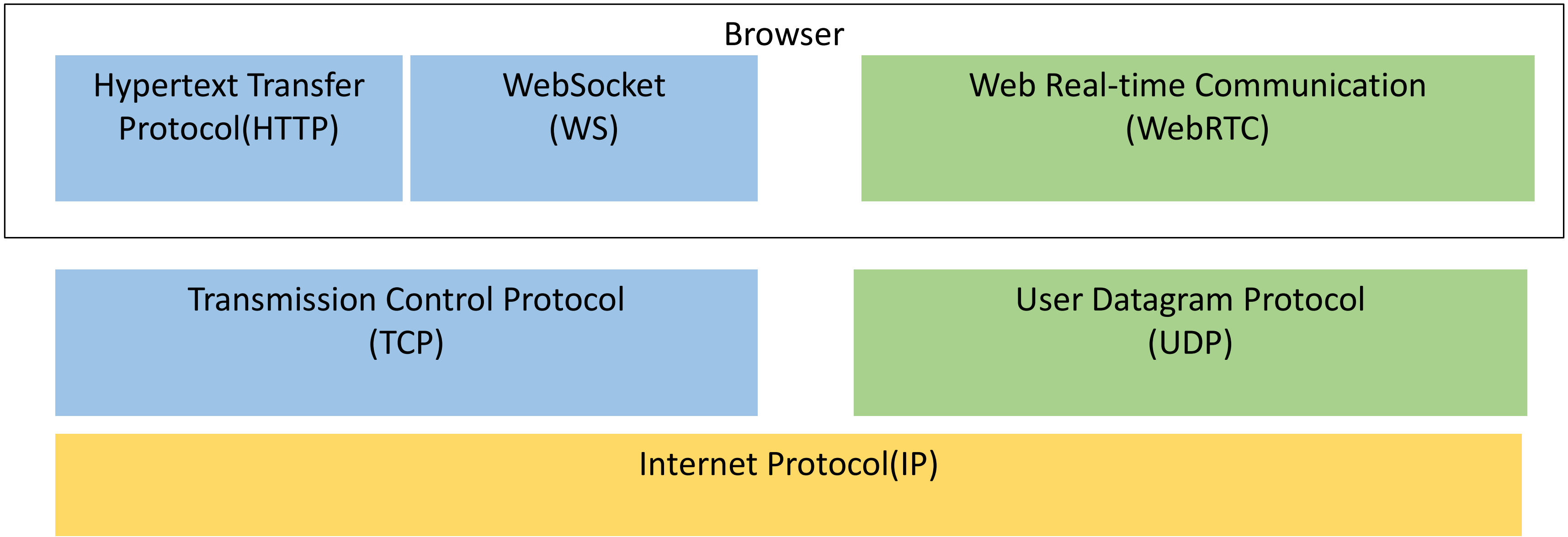
Connection between Computers

- To optimize the voice call, it depend on the connection between computers.
- TCP vs UDP

Feature	**TCP**	**UDP**
Connection	Connection-oriented (requires a handshake before data transfer).	Connectionless (no handshake, just sends).
Reliability	Reliable (guarantees delivery, retransmits lost packets, checks order).	Unreliable (no guarantee of delivery or order).
Speed	Slower (because of error checking, acknowledgments, retransmissions).	Faster (minimal overhead, no retransmission).
Data Order	Maintains order of packets.	No ordering – packets may arrive out of order.
Error Checking	Yes (with error correction).	Yes (with checksum), but no correction.
Overhead	Higher (because of headers: 20–60 bytes).	Lower (header: 8 bytes).
Use Cases	Web browsing (HTTP/HTTPS), email (SMTP/IMAP), file transfers (FTP), remote login (SSH).	Live streaming, online gaming, VoIP, DNS queries.

Web Protocols-WebSocket Vs WebRTC

- Higher level protocol on browser



Browser

Hypertext Transfer Protocol(HTTP)

WebSocket (WS)

Web Real-time Communication (WebRTC)

Transmission Control Protocol (TCP)

User Datagram Protocol (UDP)

Internet Protocol(IP)

Voice Activity Detection (VAD)

What it is:

- A fundamental process in voice AI that distinguishes speech from non-speech in an audio signal.

How it works:

- It segments audio into "speech" and "non-speech" segments, often based on detecting periods of silence or other acoustic features.

Purpose:

- It serves as the foundational component for other real-time audio processing tasks.

Agent Turn Detection (TD)

What it is:

- The higher-level process of deciding when a user's conversational "turn" has ended and the AI should respond.

How it works:

- It builds upon VAD by using additional signals and models to provide a more nuanced understanding of the conversation's flow.

Key techniques:

- **Silence-based (heuristic):** A simple method that assumes a turn ends after a set period of detected silence.
- **Context-aware/Semantic:** More advanced models that incorporate linguistic and semantic cues to predict when a user has naturally finished their thought, even if there are natural pauses within the speech.
- **Acoustic:** Uses acoustic features of speech to inform turn detection decisions.

Relationship Between VAD and Turn Detection

VAD as a foundation:

- VAD provides the essential speech-vs-non-speech information, which is a critical input for turn detection.

Turn detection for better experience:

- While VAD can identify when a person starts and stops speaking, turn detection adds a layer of "understanding" that is crucial for smooth, human-like conversations, preventing interruptions during natural pauses.

Importance in Voice Agent

- Enables real-time interaction: Both are essential for a voice agent to engage in a dynamic, back-and-forth conversation rather than operating in a disconnected, back-and-forth fashion.
- Improves user experience: Effective turn detection, particularly context-aware models, avoids abrupt interruptions and awkward silences, leading to more natural and efficient communication.
- Cost reduction: Accurate VAD can prevent processing and sending large amounts of voiceless audio to expensive speech-to-text pipelines.

AI Voice Agent Use Cases

Type	Purpose	Example Capabilities
Virtual Assistants	General-purpose agents handling diverse tasks across domains	Siri, Alexa, enterprise personal assistants
Customer Service Agents	Provide product support, troubleshooting, and escalation	Handle FAQs, detect frustration, transfer to human agent
Appointment Schedulers	Manage calendars and bookings efficiently	Schedule meetings, confirm appointments, send reminders
Information Retrievers	Deliver targeted information from internal or public data	Access knowledge bases, databases, or documents
Transactional Agents	Execute business transactions and workflows	Process payments, bookings, or orders with backend integration
Industry-Specialized Agents	Designed for specific sectors and terminology	Healthcare scheduling, financial advisory, logistics coordination

References

- https://www.voicespin.com/glossary/voice-activity-detection/
- https://www.youtube.com/watch?v=xWhI8RkRSGQ&t=60s