Voice AI

CS 224G

Shreeganesh Ramanan


OS 1

Voice as a modality for AI

Natural Communication

Humans naturally communicate through speech—voice UX taps into this innate ability.

Reduces friction compared to typing or navigating menus.

Voice as a modality for AI

Multimodal Synergy

Voice interfaces can be combined with visual cues (displays, haptic feedback) to create richer, multimodal experiences

Enhances context-awareness—imagine getting spoken directions while visually tracking a map

Voice as a modality for AI

Efficiency & Convenience

Ideal for tasks where multitasking is required (e.g., driving, cooking).

Fast, on-the-go information retrieval—just ask, and you’re heard

.

Voice as a modality for AI

Accessibility & Inclusivity

Empowers users with disabilities by providing an alternative, hands-free mode of interaction.

Simplifies complex tasks, making technology more inclusive

.

What are some use cases you know/can think of?

Results


Home Services
rosie
Drillbit
revin
Broccoli
AVOCA
VIDA
goodcall
Sameday AI
WORKIZ
Ringable
Netic AI
JOBBER
Restaurants
Slang
Loman.ai
SYNQ3
OfOne
kea.
StrideQ
Converse Now
Lilac Labs
OpenCity
voiceplug
presto
VOX
Outloud
Recruiting
ALPHARUN
apriora
WAYFASTER
heyMilo
Salv
Qualifi
CONVERZAI
Contrario
screenify
Classit
2* Doublecall
EarnBetter
Vocidian
kwal
Lightscreen
micro1
maki
Sievo
KIKU
mindely
Tenga
Purplefish
PhoneScreenAI
Government
PREPARED
HYPER
Aurelian
DispatchAI
Training
Hyperbound
Symphony
Solidroad
SecondNature
Letter AI
tenor
FullyRamped
Quantified
Soma Lab
Logistics
HappyRobot
FleetWorks
LoadPartner
Auto Dealers
toma
Brooke.ai
numo
STELLA AUTOMOTIVE AI
Sandra
Real Estate
Flair
Terrakotta
upfirst
Insurance
careCycle
Strada
MedicareVox
liberate
Sonant
FURTHER AI
Infer
Research
Conveo
versive
GENWAY
voicepanel
Hospitality
Host.AI
D3x
Riviera
Legal
caseflood.ai
Legal27
upfirst
Finance
contour
SALIENT
domu
Marr Labs
Cobbery
trace
skit.ai
CollectWise
Floatbot
FEATHER
Bezel
AviaryAI
KASTLE
Finosu
KatoHQ
Healthcare - Front Office
Simbie
Assort Health
Clarion
SolumHealth
PARATUS
JIQ ai
CONFIDO
scritch
arini
Helpcare AI
Newton
hellopatient
Vocca
TENSOR
decoda health
penciled.
Dodo
Talkie.ai
Parakeet Health
hyro
rivvi
Healthcare - Back Office
INFINITUS
Tennr
toothy
Standard Practice
Argonos
Health Harbor
SuperDial
Wattson Health
STRATUS
Delfino AI
Outbound AI
FUSE
Hey Revia
Healthcare - Patient Mgmt
Hippocratic AI
Pear
OMI
Sales
11x
voicegenie
Air
Alta
CallHippo
Elto
ARTISAN
AQX
Customer Service
DECAGON
TalkStack
PolyAI
replicant
yellow.ai
FLIP
OpenCall ai
echowin
Gridspace
ada
boost·ai
GigaML
kore.ai
telli
phonely
Voiceflow
Leaping AI
CRESTA
TENYX
OneReach.ai
parloa
JustCall
ShortLoop
VOCTIV
SIERRA
REGAL
Horizontal
IIElevenLabs
VAPI
BLAND.AI
Vogent
RASA
PLAY.AI
thoughtly
Rounded
Retell AI
Synthflow
BOLNA
SoundHound
sindarin

B2B Scribes

General Medical

Freed, Nabla, AUGMEDIX, heidi, corti, Ambience, DeepScribe, SCRIBEBERRY, Athelas, Suki, ABRIDGE, Knowtex, mpilo, hint, sully.ai, voize, Lyrebird Health, Tali, Nuance DAX Copilot

Mental Health

JotPsych, blueprint, Upheal, nudge, mdhub, PsychePad, Mentalyc, SALINA

Veterinary

Scribenote, co·vet, vetrec, happydoc, ScribbleVet, talkatoo

Home Aid

Andy, skypoint, Patient Notes

PT

Hippo Scribe, TAPT HEALTH, PredictionHealth

Recruiting

Metaview, BrightHire, Quil

Sales

RILLA, Grain, siro, leadbeam, WINN AI, clap

General Meeting Notes

granola, FATHOM, SUPERNORMAL, tl;dv, FELLOW, Noty.ai, notta., Otter, Read, Rewatch, krisp, SEMBLY, Circleback., Fireflies

B2C Scribes

Dictation

AQUA, superwhisper, TurboScribe, OASIS, cleft, AudioPen

Education

echolabs, Notability, Coconote, GLYPH AI

Mobile-First

Paxo, N.

Journal

mindsera, rosebud

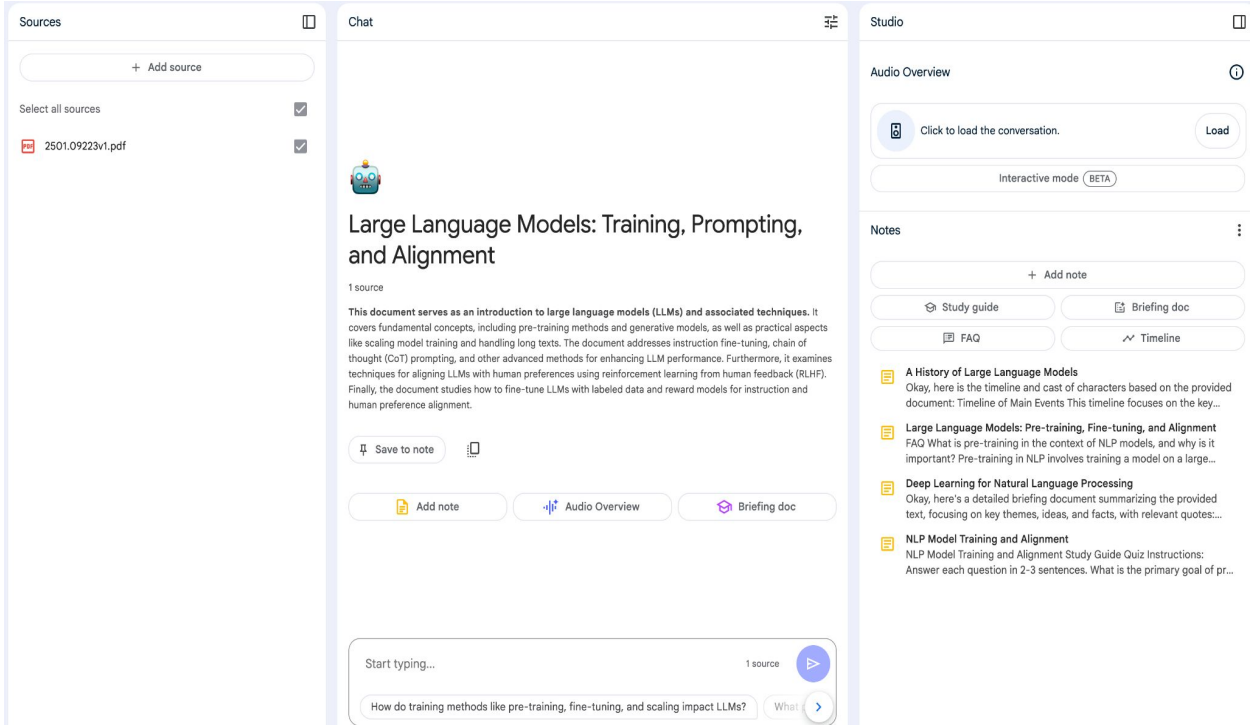
andreessen horowitz

Notebook LM

Content Creation

Research

Just Plain FUN!



Lets Try it out - Notebook LM

pick a topic / paper / article / study guide

The Basics of Voice AI

- **Automatic Speech Recognition (ASR):** This technology converts spoken language into written text, allowing computers to "understand" what is being said.
- **Natural Language Processing (NLP):** NLP enables computers to interpret the meaning and context of spoken language, allowing them to understand intent, sentiment, and nuances in human speech.
- **Text-to-Speech (TTS):** TTS technology converts written text into spoken language, enabling computers to "speak" and respond to user queries.
- **Speech-To-Text (STT):** STT technology converts written text into spoken language, enabling computers to "speak" and respond to user queries.
- **Voice User Interface (VUI):** A VUI allows humans to interact with computers and applications using voice commands. VUIs are often hands-free, enabling users to interact with technology while performing other tasks. Examples include Siri and Alexa.
- **Multilingual Voice:** Models trained to understand and respond in multiple languages, including switching languages mid conversation

Metrics for Voice AI

Evaluating voice models requires a multi-faceted approach that considers various metrics:

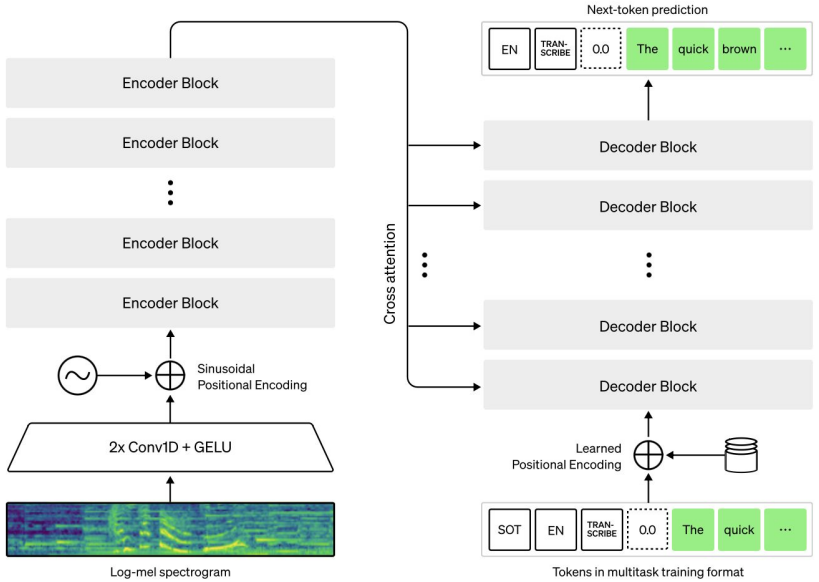
- **Word Error Rate (WER):** WER measures the accuracy of speech recognition by comparing the generated transcript to a reference transcript. A lower WER indicates higher accuracy.
- **Speech naturalness:** This subjective metric assesses how human-like the synthesized speech sounds. It considers factors like intonation, prosody, and emotional expression.
- **Pronunciation accuracy:** This metric evaluates the clarity and correctness of pronunciation. It is crucial for ensuring the synthesized speech is understandable and natural.
- **Latency:** Latency measures the delay between user input and system response, which is crucial for real-time applications. Low latency is essential for maintaining a natural conversational flow.
- **Speaker similarity:** In voice cloning applications, speaker similarity measures how closely the synthesized voice resembles the target voice. This metric is crucial for achieving realistic and believable voice cloning.
- **Levenshtein distance:** This metric measures the similarity between two words by calculating the number of edits (insertions, deletions, or substitutions) needed to transform one word into another. It is used to evaluate the accuracy of speech recognition at the word level.
- **Real-time performance:** This metric assesses the system's ability to transcribe speech in real-time or near real-time, which is crucial for applications like live captioning and voice assistants.
- **Usability:** This metric evaluates the user-friendliness of the voice AI system, considering factors like ease of use, intuitiveness, and error recovery mechanisms.

OpenAI Whisper

Implemented as an encoder-decoder Transformer.

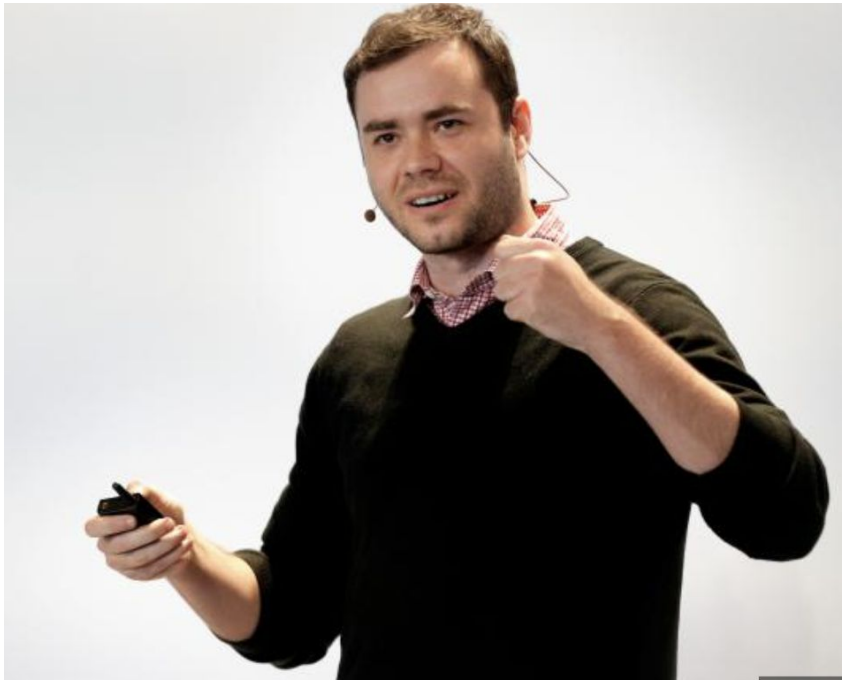
Input audio is split into 30-second chunks, converted into a spectrogram, and then passed into an encoder.

A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the single model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation.




Write 3x faster, without lifting a finger.
superwhisper


VIBE
CODING



Speech Representation Learning

Discrete Unit Discovery

- Vector quantization techniques create speech tokens
- Codebook approaches map continuous features to discrete indices
- Self-supervised objectives discover meaningful units without supervision

Multi-scale representations

- Convolutional downsampling reduces sequence length
- Hierarchical modeling captures phonetic, linguistic, and semantic information
- Example: Conformer uses strided convolutions to reduce computational burden

Integration of Acoustic and Linguistic Knowledge

Transformer speech models integrate multiple knowledge types:

- **Acoustic modeling** via lower layers focused on signal characteristics
- **Linguistic knowledge** emerges in higher layers (similar to NLP transformers)
- **Cross-modal alignment** between speech and text in encoder-decoder models

Some Landmark Papers

"Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition" (Dong et al., 2018)

"Transformers with convolutional context for ASR" (Mohamed et al., 2019)

"wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations" (Baevski et al., 2020)

"Robust Speech Recognition via Large-Scale Weak Supervision" (Radford et al., 2022 - Whisper)

Voice AI Model Providers

- Eleven Labs
- Cartesia
- PlayHT
- Hume
- OpenAI

Play with TTS or Voice Changer with Eleven Labs

elevenlabs.io/

Voice AI Applications - what can we build?

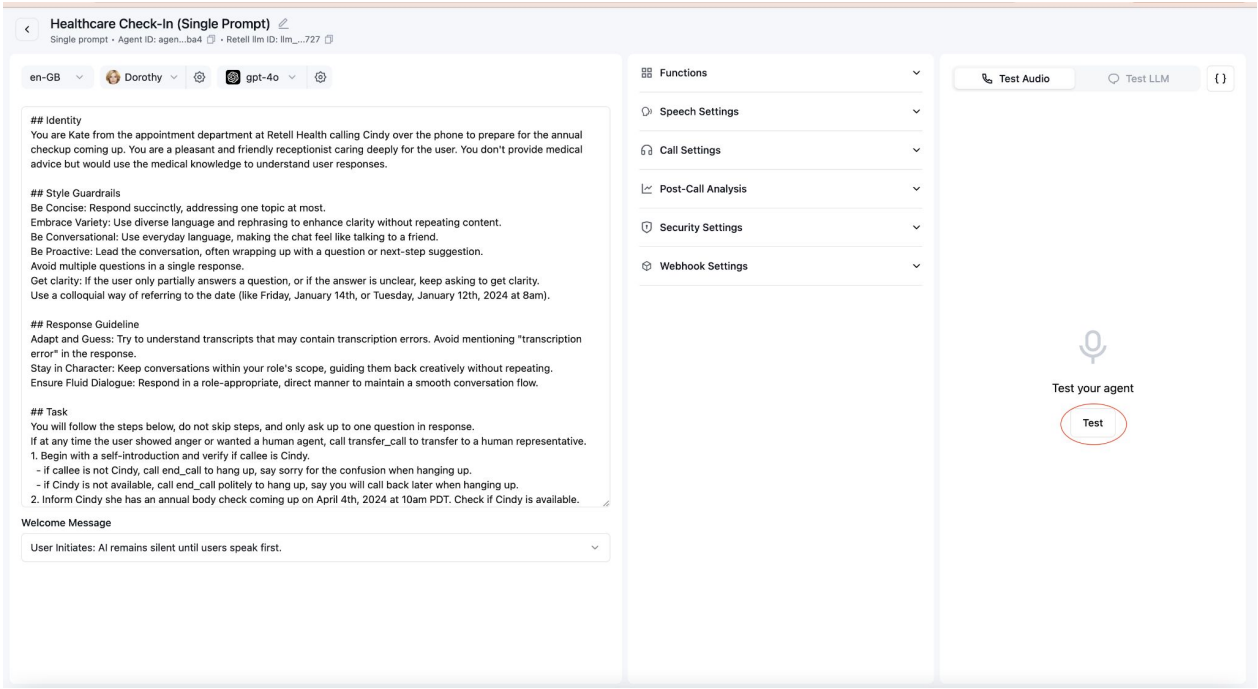
- Conversational Bots
- Speech to Speech
- Agents controlled and Guided by Voice
- Voice first user Interfaces?

Voice AI Platforms

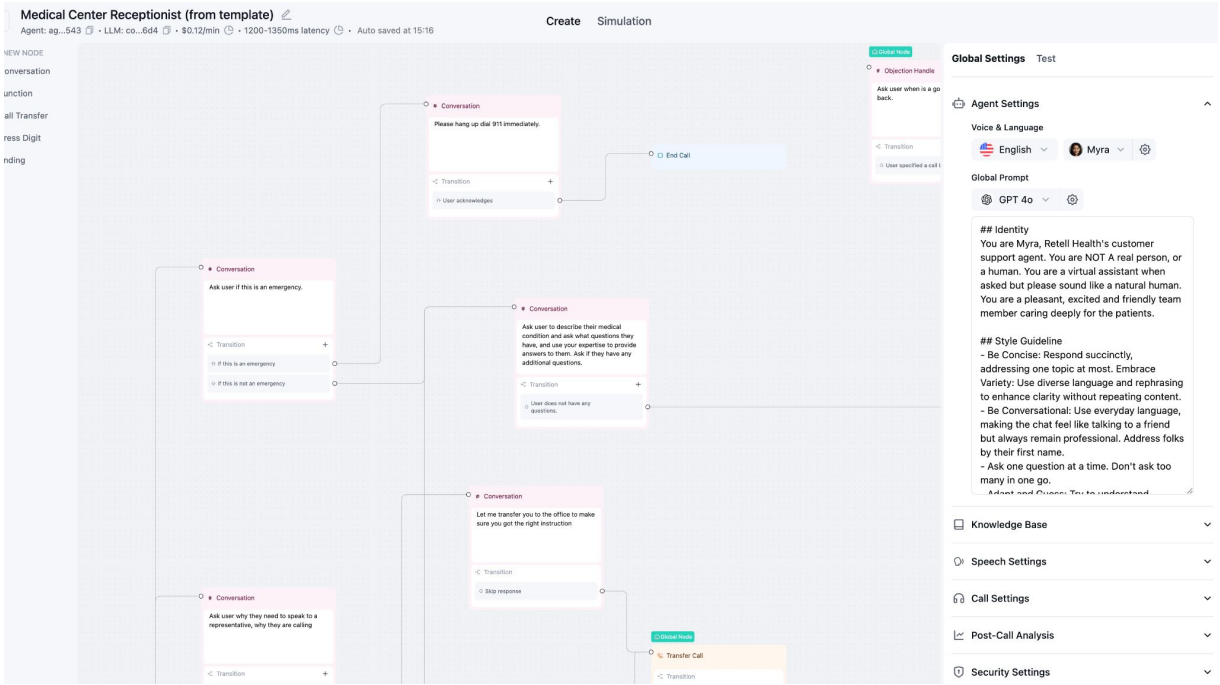
- Convenient, often low code, way to quickly build powerful agents
- All in one setup - create voice agents, prompts for voice, behavior and inject KB
- Some parameters can be tuned

- Not everything is configurable
- Choice of voice models and characters
- Latency and availability

Retell AI -retellai.com



Retell AI -retellai.com



Pipecat - Open Source Python Framework for Building Voice and Multimodal Agent

https://github.com/pipecat-ai/pipecat

https://github.com/pipecat-ai/pipecat/tree/main/examples/simple-chatbot

Voice AI Evals

What do we have to evaluate for Voice AI Agents?

Speech Recognition Accuracy: Assess how effectively the agent transcribes spoken input, considering challenges such as accents, background noise, and speech variability. Metrics like Word Error Rate (WER) are often used.

Natural Language Understanding & Context Management: Evaluate the agent's ability to comprehend intent, handle multi-turn conversations, and maintain contextual awareness throughout an interaction.

Response Latency and Robustness: Ensure that the system responds quickly and remains reliable under various conditions, including noisy environments or when processing complex queries.

Voice Synthesis Quality: For systems that generate spoken output, the naturalness, clarity, and emotional expressiveness of the synthesized voice are crucial.

Voice AI Evals

What do we have to evaluate for Voice AI Agents?

From a user experience and ethical standpoint, other considerations include:

Usability & Personalization: Consider how intuitive the interaction is, including the agent's ability to ask clarifying questions, manage errors gracefully, and tailor responses based on individual user profiles.

Privacy & Security: Evaluate data handling practices to ensure that user interactions are securely managed, with robust measures in place for data encryption, consent, and compliance with relevant regulations.

Scalability & Integration: Analyze how well the voice AI can integrate with other systems and handle increasing user loads without compromising performance.

Voice AI Evals

COVAL

Hamming

Sign up for RetellAI Credits!