

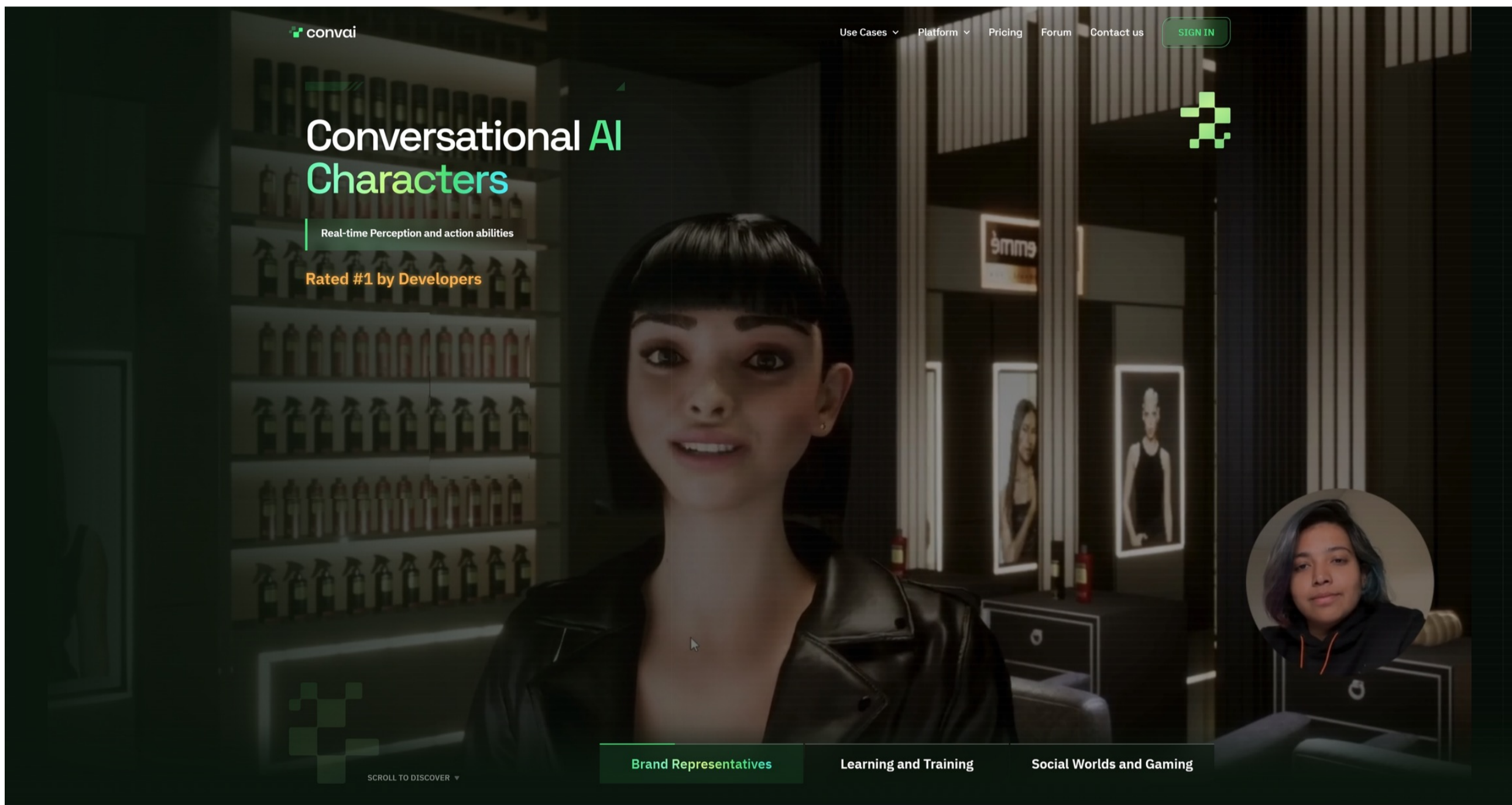
Towards Human–Avatar Communication

Training Hearing Impaired via Emotionally Adaptive Voice Synthesis

Background: LLM-driven avatars are increasingly employed in VR for training. How about Hearing-impaired?

Bottleneck: Text-to-speech (TTS) : high latency, limited expressiveness, and lack of distinct speaker identities.

Can our real-time voice conversion plug-in independently control emotional sentiment and vocal timbre?



Select Language and Speech [Beta] **SoTA: convai**
[Learn more on how to use Voice and Languages](#)

Set Language

Add Custom Pronunciation

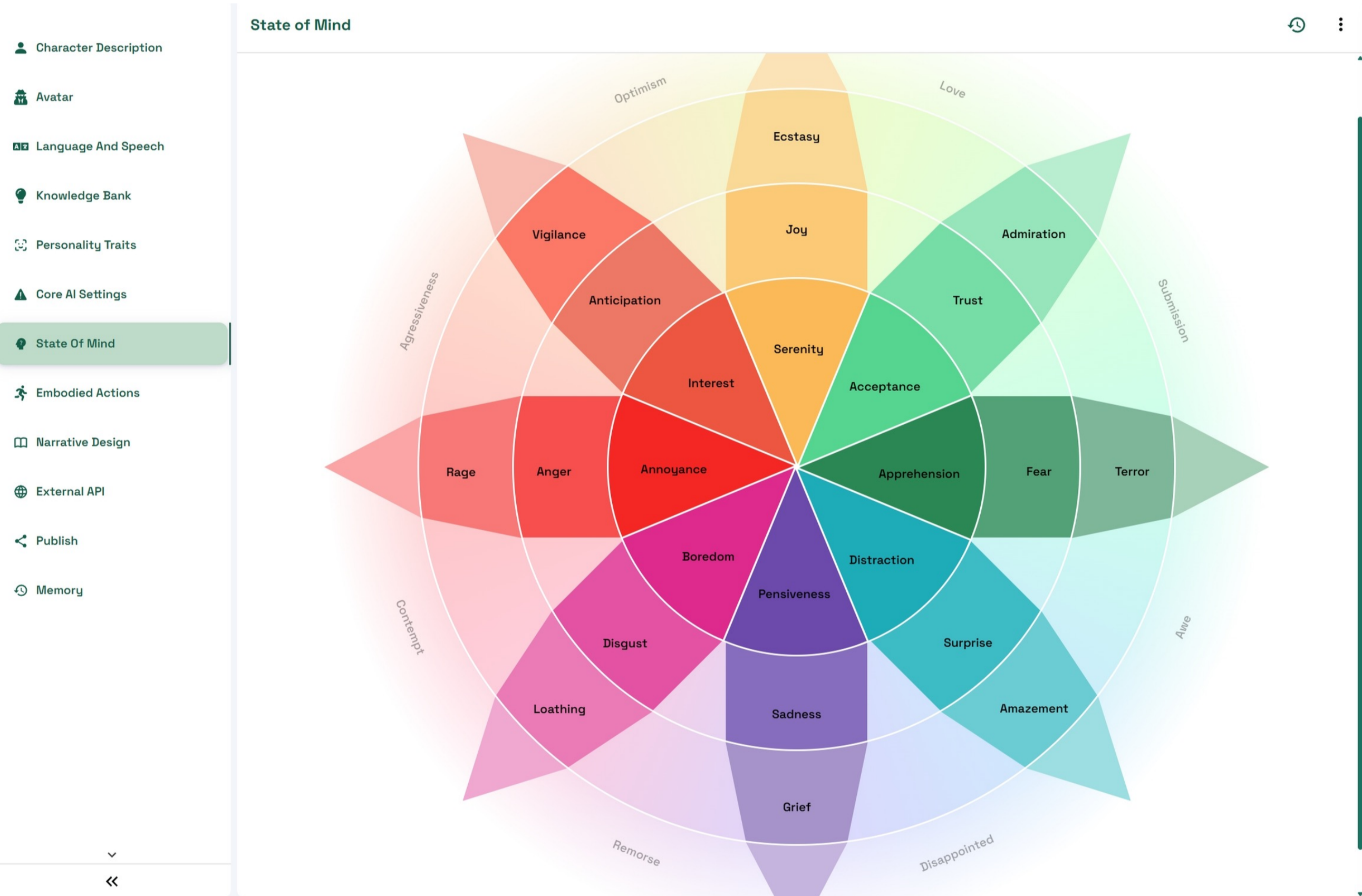
New Word Recognition

Languages [Select up to 4 languages]

English × Select a language

Voice

Daniel (Young American Male Voice)



Text To Speech and Speech To Speech

Is Recording
Target is Convai Player
Target self Return Value

Start Recording
Target is Convai Player
Target self

Finish Recording
Target is Convai Player
Target self Return Value

Is Talking
Target is Convai Player
Target self Return Value

Send Text
Target is Convai Player
Target self Convai Chatbot Component Text Environment Select Asset Generate Actions Voice Response Run on Server

Start Talking
Target is Convai Player
Target self Convai Chatbot Component Environment Select Asset Generate Actions Voice Response Run on Server Stream Player Mic

Finish Talking
Target is Convai Player
Target self

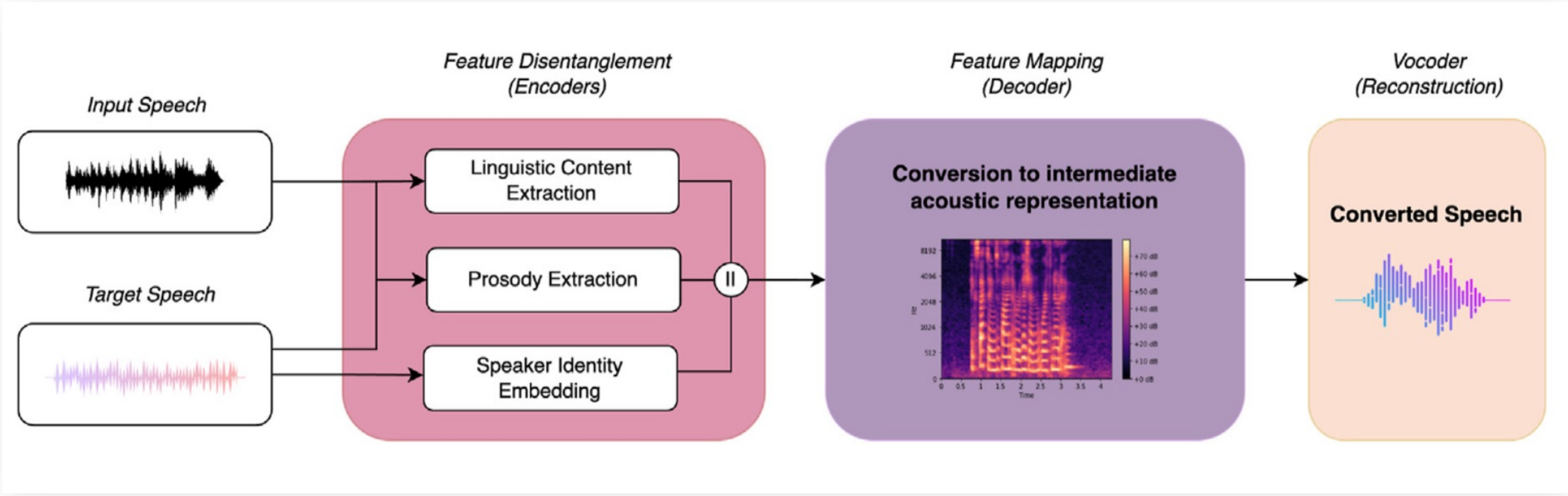
Microphone settings

Get Available Capture Device Names
Target is Convai Player
Target self Return Value

Set Capture Device by Name
Target is Convai Player
Target self Device Name Return Value

Get Active Capture Device
Target is Convai Player
Target self Out Info

Our work Plugin Post TTS inference Extensive Disentanglement Voice Cloning (20 sec)



Aim: Test the recognition of emotional states and speaker variations in VR.



AALBORG
UNIVERSITET

Anders R Bargum, Stefania Serafin, Cumhur Erkut
cuerlab
Copenhagen, Denmark

