Use Case: Real-time ASR handling of code-switching…

ID	Name	Description
UL0	Real-time text output	As user actively speaking at a normal rate (100 words per minute), shall return translations in a comparable amount of time compared with typing
UL1	Cross-device support	Our app shall be usable on different laptop devices, laptop OS
UL2	Reasonable output	Shall not display gibberish; shall recognize proper vocabs in English and Mandarin
UL3	Noise tolerance	Shall remain some accuracy when there is noise in the audio
UL4	Reasonable speech length	Shall have the capacity to transcribe a useful quantity of speech in one shot
UL5	Reasonable vocab recognition	Shall recognize vocabularies used in daily life conversations in English and Chinese

Design Requirements

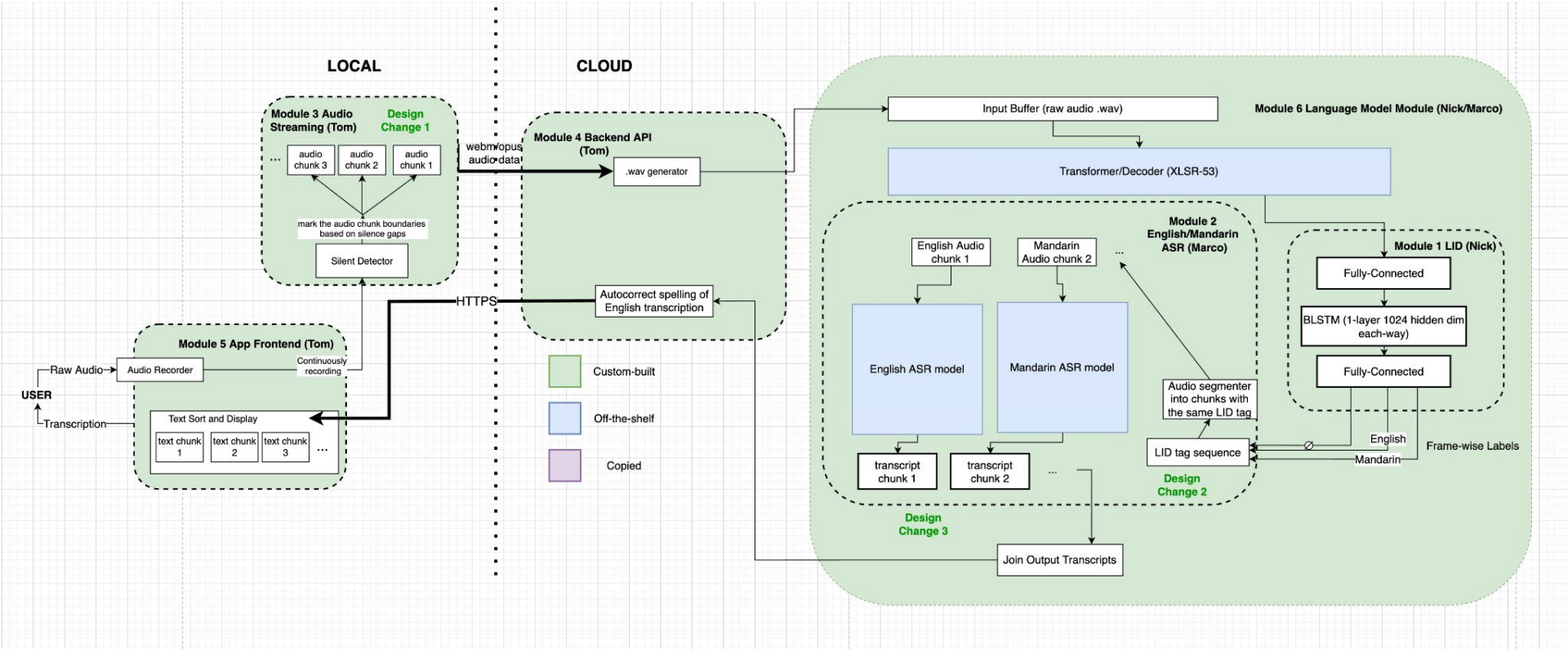
ID	Name	Description	Value	Test	UL Req.
DR0	End-to-end Latency	Shall return a translation of the first segment in no more than 1 second	1s	ST0	UL0
DR1	Throughput	Shall process audio faster than audio is inputting, lower than 1.0 s/s	1.0 s/s	ST1	UL0
DR2	Character Error Rate	Shall transcribe speech at a translation error rate of no more than 25%	25%	ST2	UL2, UL5
DR3	Noise Tolerance	Shall remain below 30% translation error rate when signal to noise ratio in the audio is lower than 20dB	20dB	ST3	UL3
DR5	Supporting Environment	Shall be usable on a laptop	N/A	ST5	UL1
DR6	Max Data per Input	Shall be able to transcribed a maximum of 1 minute of audio in a single input instance from a user	1 min.	ST1	UL4

Current Design

6 Modules, Independently Testable

1. **Language Identification (LID):** Classifies frames into Eng., Mand., Blank
2. **Automatic Speech Recognition (ASR):** Single language ASR model transcribes each audio sequence based on its LID tags
3. **Audio Streaming:** Detect silence gaps within the ongoing recording and calls remote API to transcribe the newest chunk
4. **Backend APIs:** pre-loads model on startup, receives audio chunks, requests chunk transcriptions from model, autocorrect English sequence within predicted transcript
5. **App Frontend:** Simple, real-time I/O interface for user to record messages and receive transcriptions
6. **Language Model:** Encapsulates and combines submodels, contains all end-to-end training and validation

Architecture



Complete Solution

English Demo:

Model struggles to detect high-level contextual info to override difficult pronunciation, hasn't seen enough of this specific connection

Text:
do you think someday people can immigrate to Mars or other planets

Prediction:
do you think someday people can emigrate to marshal other planets

Complete Solution

Mandarin Demo:

Translation:
yes i feel like they cooperate very well

Text:
对我感觉他们配合的时候默契特别好

Prediction:
对我我感觉他们配盒的时候默契特别好

Complete Solution

Code-switch Demo:

Translation:
my major is electrical and computer engineering

Text:
我的专业是electrical and computer engineering

Prediction:
我的专业是electrical and computer engineering

Complete Solution

Code-switch Demo:

Translation:
besides if I really wanted to hide from him on purpose
why would I even ask him at this time
I can wait until the interview is over and then tell him

Text:
再说我要是真的想故意瞒着他 why would I even ask him at this time 我可以到interview全部结束了再 tell him

Prediction:
在说我要是真的想故意蛮着它 why would i even ask him at this time 我可到or interview全部结束了a tell him

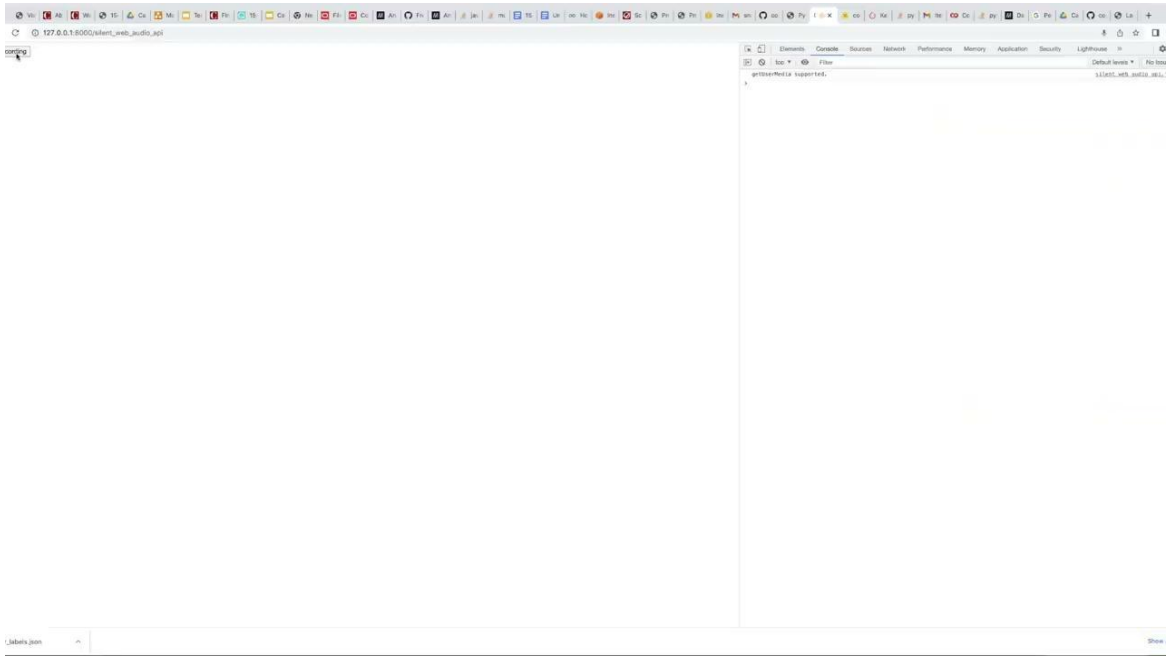
Complete Solution

Real-time Demo:

Reference: 我现在来测一下中文大概的样子, and I am going to speak English now. And I had salad and eggs for breakfast today.

Translation: I am going to test Chinese, and I am going to speak English now. And I had salad and eggs for breakfast today.

Prediction: 我现在来测一下中文大概的样子 and I’m going to speakenglish now 埃 ey had saladandeggs for breakfasttoday.



Test Results

Name	Description	Testing Procedure	Results
End-to-end Latency	Shall return a translation of the first segment in no more than 1 second	Testing the time difference between request to AWS instance and returned result	On average 600 ms
Throughput	Shall process audio faster than audio is inputting, lower than 1.0 s/s	Testing average processing time measured between input and output on AWS instance	0.7s / s (takes 0.7s for server to process 1s of audio)
Character Error Rate	Shall transcribe speech at a translation error rate of no more than 25%	Testing CER on various datasets	22.94% on code-switch dataset 18.25% on dataset containing English, Mandarin, and code-switch speech
Noise Tolerance	Shall remain below 30% character error rate when signal to noise ratio in the audio is lower than 20dB	Testing on dataset with low signal to noise ratio	31.28%
Supporting Environment	Shall be usable on a laptop	Running on Python local environment and through Chrome browser on MacOS and Windows	Local python script is able to run on any OS Interactive mode is able to run with Chrome browser
Max Data per Input	Shall be able to transcribed a maximum of 1 minute of audio in a single input instance from a user	Testing single long segment of speech longer than 1 minute	Is able to support 1 minute audio However, GPU memory puts a cap on the maximum single segment ~10s

Trade-offs

Audio Chunking

- Larger chunk strongly improves accuracy → longer latency
  - **Soln: Automatic chunking on large silence balances latency and accuracy**

Model Size

- Single encoder **~1.2 GB**, baseline for performance → unusable accuracy
  - **Soln: separate ASR and LID models**
- Separate encoders **~2.4 GB**, flexible architecture → size makes mobile deployment less feasible

LID-ASR Fusion Techniques

- *Jointly trained combined model*
  - Difficult to label audio segments for LID on the fly → unstable training
  - Complex training, decaying hyperparameters needed → optimal interpolation between LID and ASR loss difficult to find
- *Multiplexed ASR*
  - LID model segments and labels audio, used to choose language specific ASR model
  - ASR model generates output transcription chunks
  - Allows for parallelization between ASR models when predicting → better latency + throughput

Comparative Accuracy

Arch.	CER*	WER*
Single_v1	0.36	1
Single_v2	0.23	0.64
Combined_v1	0.23	0.59
Combined_v2	0.22	0.58
Muxed_v1	0.22	x

Hyperparameter Tuning

λ	0	0.1	0.2	0.3
Best WER*	1.00	0.74	0.58	0.66

**Performance on total eval set*

Schedule Updates