Boston University
Electrical & Computer Engineering

Boston University
Electrical & Computer Engineering
EC 464 Senior Design Project

Second Prototype Test Report

BUtLAR

By
Team 12
Digital Human - Yobe

Team Members

Noa Margolin noam@bu.edu
Suhani Mitra suhanim@bu.edu
Jackie Salamy jesalamy@bu.edu
Andrew Sasamori sasamori@bu.edu

Required Materials:

Hardware:

- Raspberry Pi V5
- Two Røde Microphones
- LCD Screen (PHO 113 computer lab monitor)

Software:

- Shell Script
    - g++
    - Python Virtual Environment
- Live Audio Processing
    - *miniaudio_stream.c*
        - Utilizes C++ miniaudio library to capture audio in live time
- Yobe SDK (GrandE)
- Audio Generation
    - *IDListener_demo.cpp*
- Context-Specific Database
    - *school.db* (two-table database with one about Photonics-based professors and course-specific information)
- Google ASR Speech-To-Text API, LLM API
    - *main.py*
        - Call the files below
    - *voiceAssistant.py*
        - Processes streamed audio and prepares for LLM processing
        - Performs API calls
    - *getAsssitance.py*
        - Corrects last names, generates SQL queries, and provides LLM responses

Second Prototype Goal: support a full conversation, employing live audio processing and low latency LLM-generated relevant responses.

Setup:

Our system setup begins with the hardware components: a Raspberry Pi connected via Ethernet to host the software on a Linux machine and two Rode Microphones for capturing audio input. The microphones are set at a standard of 9 inches apart, facing upward. The pipeline is driven by a Bash script that automates the processes of audio capture, processing, and response generation. As depicted in Figure 1, the backend workflow captures audio processes using Yobe's SDK in live time. The pipeline then performs speech-to-text transcription, once again in live time. Once the full question is processed, the OpenAI-powered LLM generates a response and utilizes our prompt engineering document for use-case-specific instances. For this test, we draw from a BU-specific database with information about certain professors' classes taught. Finally, the LLM-generated response is conveyed through a digital human, enabling seamless and interactive UI engagement.

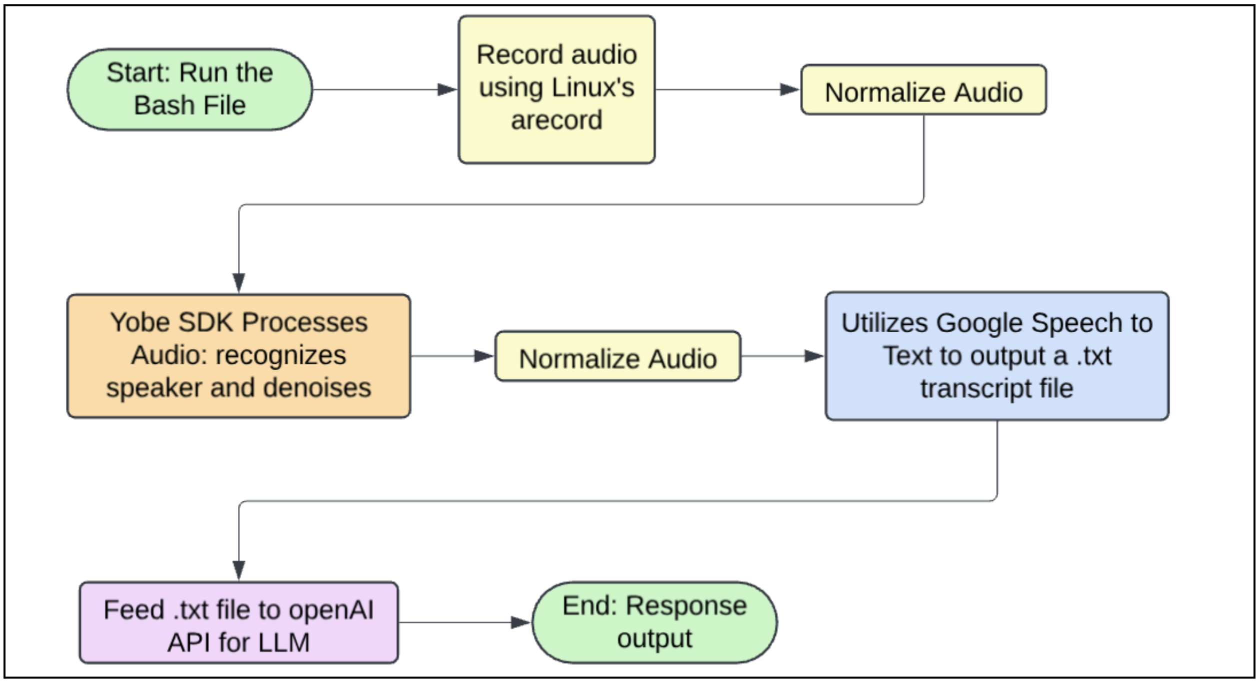

Start: Run the Bash File
Record audio using Linux's arecord
Normalize Audio
Yobe SDK Processes Audio: recognizes speaker and denoises
Normalize Audio
Utilizes Google Speech to Text to output a .txt transcript file
Feed .txt file to openAI API for LLM
End: Response output

Figure 1: Illustration of Backend System Integration

Pre-Testing Setup Procedure:

Raspberry Pi Connection:

1. 2 AI-Micro Rode Dual Speakers are connected to Raspberry Pi.
2. Raspberry Pi is connected to the network via Ethernet.
3. Run and pipe the live audio streaming files on the Raspberry Pi.

Server-Side Connection:

1. Establish SSH connectivity with the Raspberry Pi (remote access) using the following command: **ssh yobe@128.197.180.176**
2. Navigate to the appropriate directory:
    a. cd BUtLAR_Voice-Powered-Digital_Human_Assistant/Audio/testing_audio

Running the Session

1. Start a GCloud virtual environment:
    a. source ~/gcloudenv/bin/activate
2. Begin live-streaming audio
    a. **./miniaudio_stream | sox -t raw -r 16000 -e signed -b 16 -c 2 - -t raw -r 16000 -e signed -b 16 -c 1 - | python3 main.py**

Testing Procedure:

8 specific tests must be evaluated as either "Pass" or "Fail." To achieve a "Pass," each test must meet its unique criteria, ensure a latency of less than 5 seconds from the end of the audio recording to transcript generation, and produce a transcript that accurately conveys the intended message.

1. Live Audio Processing
    a. Intakes audio in live time
2. Name Correction
    a. Mispronounced names with inaccurate transcriptions are matched and output the correct last name
    b. Last name is properly detected from the database list
3. Conversation Termination
    a. Query thread ends when the user says, "Goodbye, BUtLAR."
4. Multiple Queries
    a. Users can ask multiple questions until they terminate the session.

Measurable Criteria:

Specific Test Case Requirements:

I. **Live Audio**
   A. Speaker can ask questions in real-time.
   B. We will say the sentence "Where is Professor Pisano's office room?" and check the latency.

II. **Name Detection**
   A. The transcript after Speech-To-Text will be checked for conveying a Professor's correct last name.
   B. We will say, "Where is Professor Eagle's office room?" → Correct Name: Egele

III. **Loop Termination**
   A. The conversation ends with the user saying, "Goodbye, BUtLAR."

IV. **Multiple Queries**
   A. Users can ask multiple questions in the same conversation thread.
   B. For our first trial, we will ask, "Who teaches Control Systems?" then ask, "Where is the course EC 471 located?"
   C. For our second trial, we will ask, "What professor teaches Control Systems?" then ask, "What is the location of the course EC 531?"

General Requirements:
In addition to satisfying the criteria above, the system must meet the following overarching requirements for every test case:

- **Latency:** The time from the end of the audio recording to the generation of the LLM-generated response must be less than 5 seconds.
- **Message Accuracy:** The transcript must accurately convey the intended message query.
- **Response Relevancy:** All answers provided must be accurate and relevant to the user's question.

Score Sheet:

Requirement	**Transcript is correct (Y/N)**	**Latency**	**Pass/Fail**
Live Audio Processing Test 1	N/A	< 2 s: 1.926	P
Live Audio Processing Test 2	N/A	< 2 s: 1.403	P
Name Correction Test 1	(Print both wrong and corrected)	< 2 s: 1.626	P
Name Correction Test 2	(Print both wrong and corrected)	< 2 s: 1.757	P
Conversation Terminates Test 1	N/A	N/A	P
Conversation Terminates Test 2	N/A	N/A	P
Asking Multiple Queries Test 1	N/A	N/A *Q1*: 1.616 *Q2*: 1.692	P
Asking Multiple Queries Test 2	N/A	N/A Q1: 1.488 Q2: 2.252*	P
Result →		8/8	

Test Data Conclusions

Our second prototype test consisted of 4 tests, each with 2 trials. The main aspect we wanted to see growth in was our latency compared to our first prototype testing. In this testing, we aimed for latency under < 2 seconds per response.

In Live Audio Processing Test 1, we asked BUtLAR, “Where is Professor Pisano’s office room?” which had a latency of 1.926s. Our second trial of the test resulted in a latency of 1.403s. These tests were run to identify whether or not BUtLAR was intaking audio in live time, which was proven to be true based on the low latency response and timely audio capture. Both trials also resulted in an accurate response from the LLM and successful SQL queries: “Professor Pisano's office room is 522.”

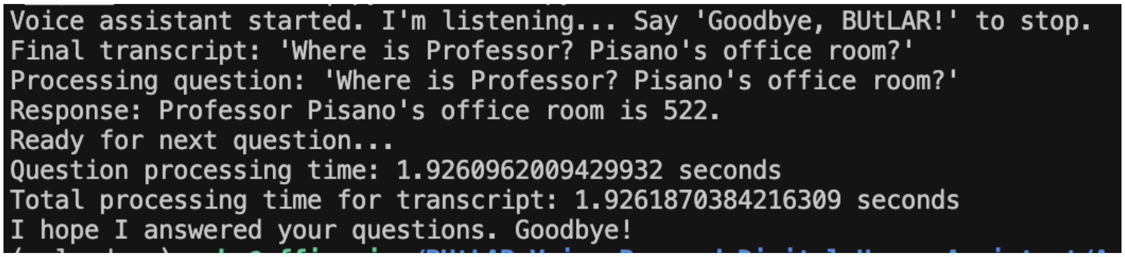
In Name Correction Test 1, we asked BUtLAR, “Where is Professor Egele’s office room?” to analyze the transcript output before and after correction. We wanted BUtLAR to have the capability to correct names if they were transcribed incorrectly or mispronounced. Therefore, in the example with Professor Egele, the transcript successfully corrected the transcription from Eagle (the user’s audio transcribed) to Egele (the correct name listed in the database). From the first trial, the latency resulted in 1.626s, while for the second trial, the latency resulted in 1.757s. Both trials resulted in a generated response of “Professor Egele's office room is 337.” Thus, these two trails successfully passed as they were under 2 seconds and corrected the mistranscribed name.

In Conversation Terminates Test 1, we tested if the BUtLAR conversation successfully terminated when the user said, “Goodbye, BUtLAR.” Conversation sessions in both trials ended upon hearing “Goodbye, BUtLAR.” *Figure 1* and *Figure 2* display examples of BUtLAR’s response when the user initiated a conversation end. This helped us identify that this phrase is consistent and can serve as a standard user-initiated ending for conversations.

Finally, in Asking Multiple Queries Test 1, we tested two questions to demonstrate that the user can ask multiple questions before terminating the session. The first and second questions we tested were “Who teaches control systems?” and “Where is the course EC 471 located?” The first trial had a latency of 1.616s for the first question and 1.692s for the second question. The responses produced were “The course ‘Control Systems’ is taught by Pisano” and “The course EC 471 is located in CAS 227” for the respective questions. For the second trial, we asked two questions of a similar format to the previous trial, just using different wording or course numbers. The first question had a latency of 1.488s, and the second question had a latency of 2.252s. The responses produced were “The professor who teaches control systems is Professor Pisano,” and “The location of the course EC 531 is CDS 264,” respectively. Though our second query resulted in a response time greater than 2 seconds, the running time for the two-question query test was under 4 seconds, which averaged each question to be under 4 seconds. We classified this test as a *Pass* as well. We concluded that, oftentimes, subsequent queries take slightly longer to render than the initial one. The results from both of these trials are detailed in *Figure 3*.

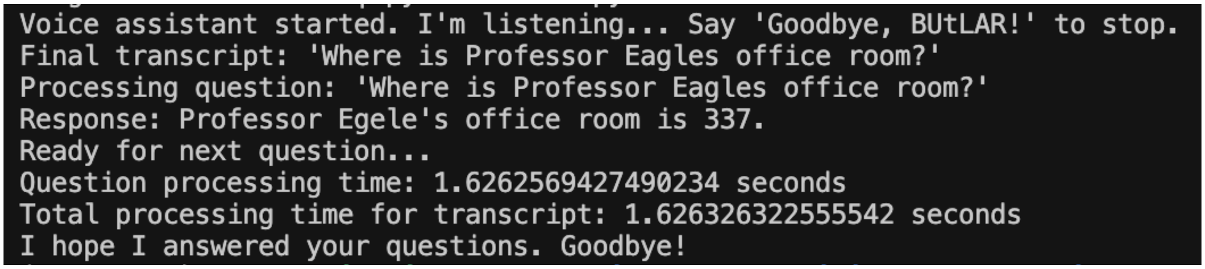
After completing all the tests and their respective trials, we concluded that all tests passed as the session operated as expected with accurate responses, and latencies were under 2 seconds. With these successes, although the latency was reduced by about 6 times, there are still aspects we must improve upon. At times, the latency is longer than expected, which is due to LLM overhead, however, we want to improve on this for the final design. Latency is our primary concern. Lower latency will simulate natural human-like conversation. There is sometimes an SQL query error if the question is asked with imperfect wording. During testing, questions that had explicit answers in the database were always asked.. Lastly, as seen in the Appendix, the output transcription (from Google ASR) often appended sudden punctuation, usually question marks and commas, which split up sentences prematurely. Due to some inconsistencies with the transcription through Google ASR, we will continue testing other Speech-to-Text platforms. Reliable transcriptions are one of our main goals when finalizing our product, as is with best response generation. In the following weeks, we will continue to tweak certain factors to obtain the highest yield in a polished LLM Auditory Responder.

Appendix



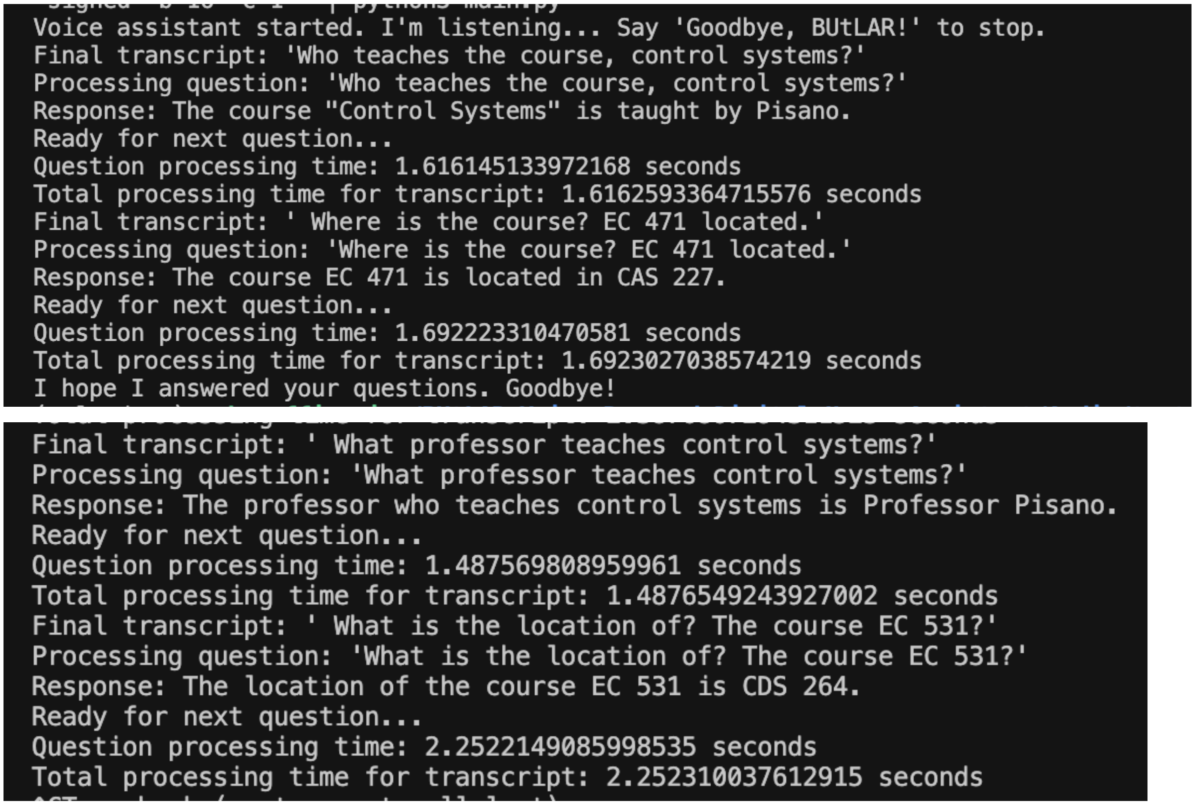
```
Voice assistant started. I'm listening... Say 'Goodbye, BUtLAR!' to stop.
Final transcript: 'Where is Professor? Pisano's office room?'
Processing question: 'Where is Professor? Pisano's office room?'
Response: Professor Pisano's office room is 522.
Ready for next question...
Question processing time: 1.9260962009429932 seconds
Total processing time for transcript: 1.9261870384216309 seconds
I hope I answered your questions. Goodbye!
```

Figure 1: Live Audio Test



```
Voice assistant started. I'm listening... Say 'Goodbye, BUtLAR!' to stop.
Final transcript: 'Where is Professor Eagles office room?'
Processing question: 'Where is Professor Eagles office room?'
Response: Professor Egele's office room is 337.
Ready for next question...
Question processing time: 1.6262569427490234 seconds
Total processing time for transcript: 1.626326322555542 seconds
I hope I answered your questions. Goodbye!
```

Figure 2: Name Correction Test



```
Voice assistant started. I'm listening... Say 'Goodbye, BUtLAR!' to stop.
Final transcript: 'Who teaches the course, control systems?'
Processing question: 'Who teaches the course, control systems?'
Response: The course "Control Systems" is taught by Pisano.
Ready for next question...
Question processing time: 1.616145133972168 seconds
Total processing time for transcript: 1.6162593364715576 seconds
Final transcript: ' Where is the course? EC 471 located.'
Processing question: 'Where is the course? EC 471 located.'
Response: The course EC 471 is located in CAS 227.
Ready for next question...
Question processing time: 1.692223310470581 seconds
Total processing time for transcript: 1.6923027038574219 seconds
I hope I answered your questions. Goodbye!
```

```
Final transcript: ' What professor teaches control systems?'
Processing question: 'What professor teaches control systems?'
Response: The professor who teaches control systems is Professor Pisano.
Ready for next question...
Question processing time: 1.487569808959961 seconds
Total processing time for transcript: 1.4876549243927002 seconds
Final transcript: ' What is the location of? The course EC 531?'
Processing question: 'What is the location of? The course EC 531?'
Response: The location of the course EC 531 is CDS 264.
Ready for next question...
Question processing time: 2.2522149085998535 seconds
Total processing time for transcript: 2.252310037612915 seconds
```

Figure 3: Multiple Queries (Trial 1 and Trial 2)