Design Review: StenoPhone

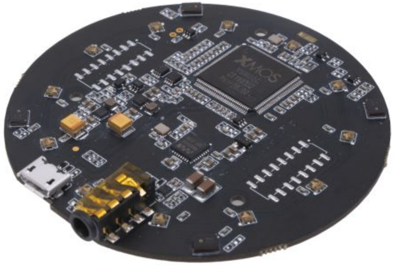
D6: Ellen Seeser, Cambrea Earley, and Mitchell Yang

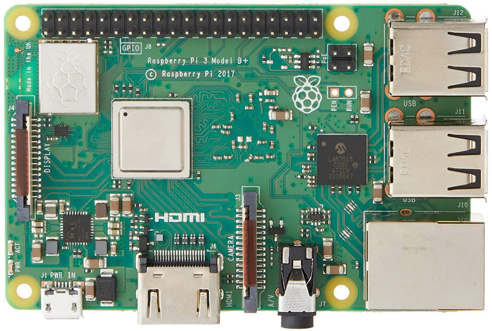
StenoPhone Recap

- Application area: distributed team meetings
- Automatic transcription ↔ recordkeeping, communication
- Meetings with multiple participants, multiple rooms

Solution Approach: Hardware

- ReSpeaker Mic Array v2.0
- 5m radius speakers
- Raspberry Pi 3 B
- SciPy, mp3 compression





Solution Approach: Software

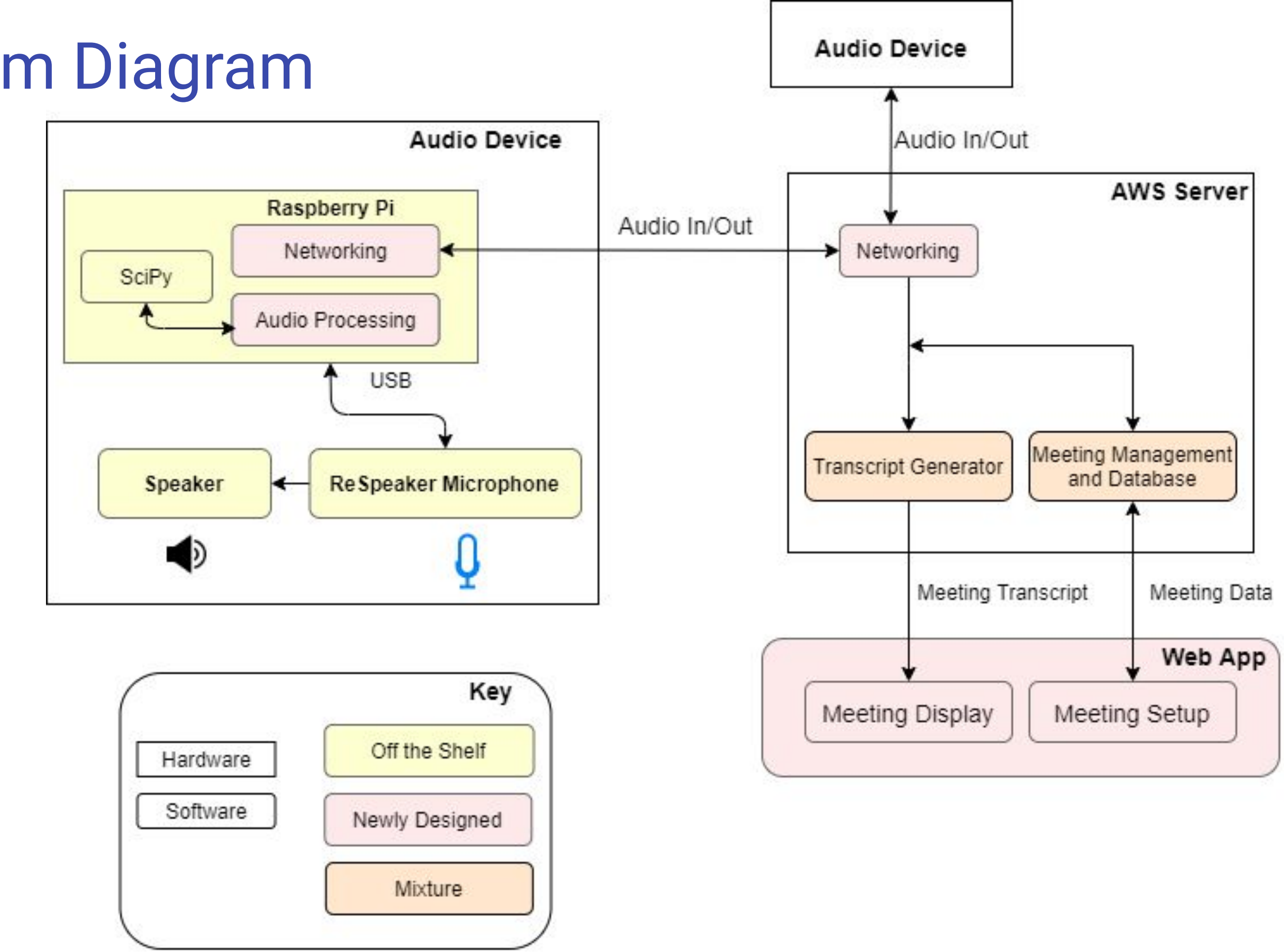
- AWS server
- Python, Django, SQLite
- Meeting management

Solution Approach: ML

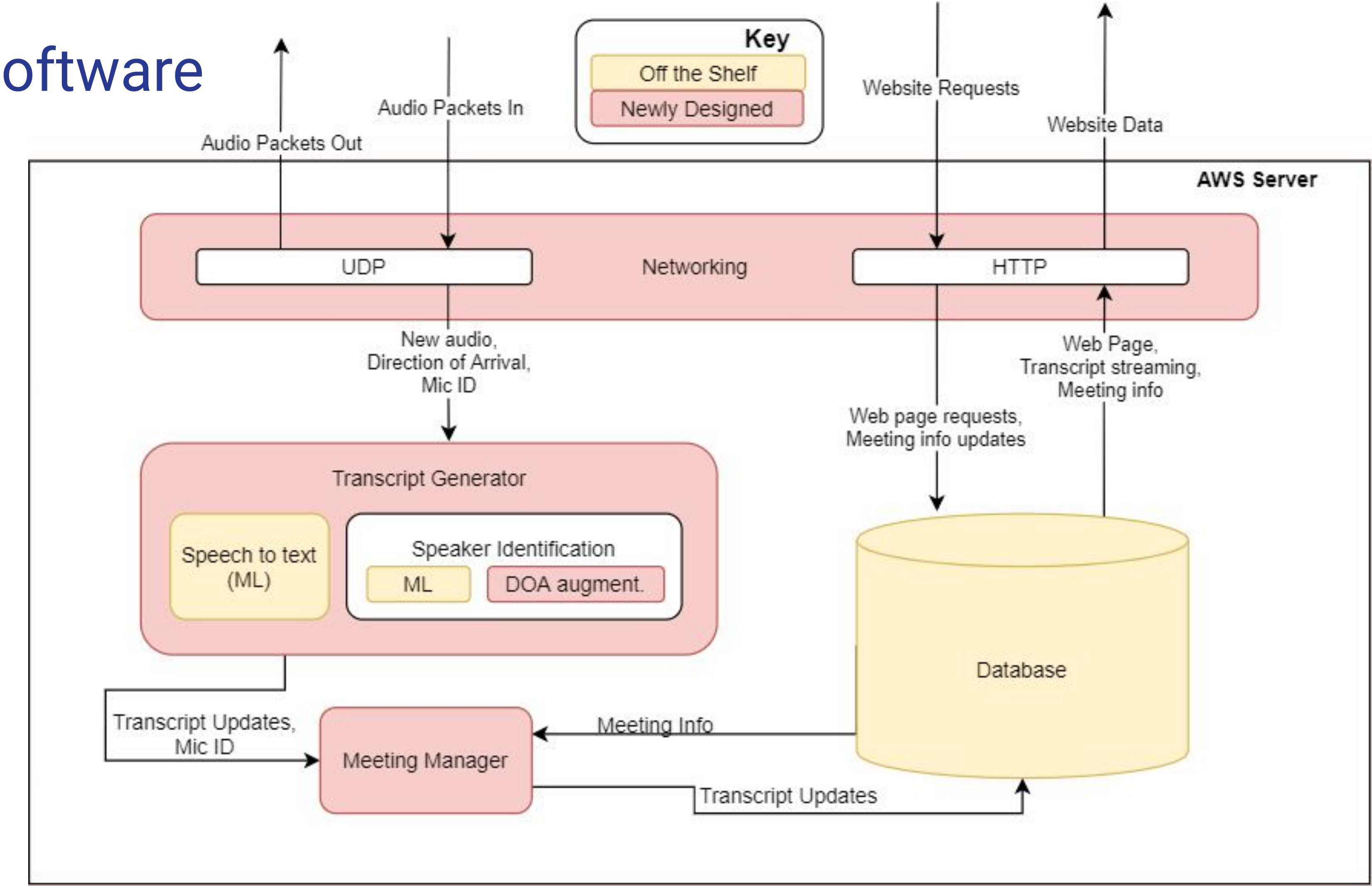
- Speech to text
  - Candidates: Mozilla DeepSpeech, Google Cloud Speech to Text

- Speaker Identification
  - ML component candidates: pyannote audio, pyBK, Hitachi Speech EEND, BUTSpeechFIT VBx
  - Direction of Arrival augmentation

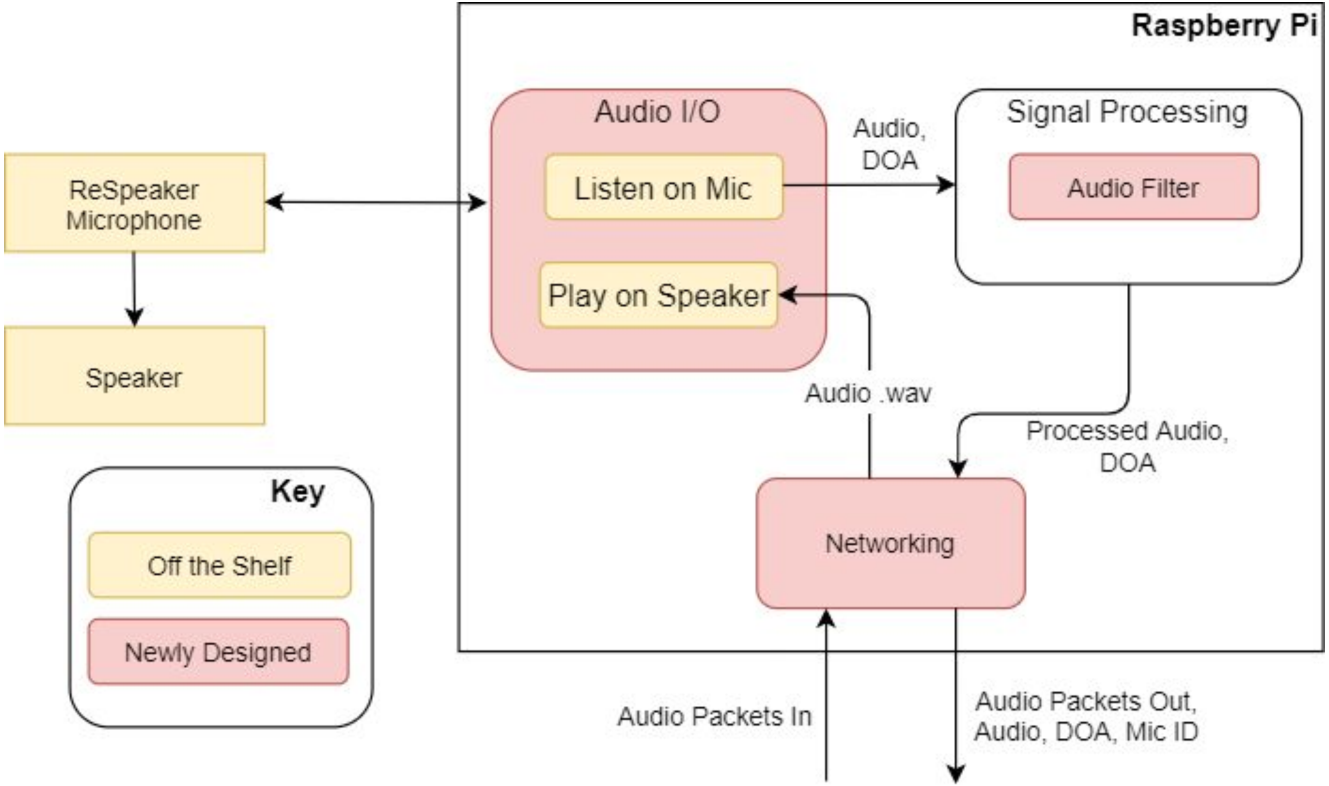
System Diagram



AWS Software



RPi Software



Metrics and Validation

Requirement	Metric	Test	Failure Remediation
Audio Transmission Latency	Mouth-to-Ear Latency (ms) < 150 ms	Route audio packets through server, to and from same microphone for timestamp comparison	Purchase better AWS specs Relocate audio processing
Transcript Latency	Average Word Delay (s) < 3s	Capture timestamp of audio captured by mic and timestamp of packet arrival in browser	Purchase better AWS specs Relocate audio or ML processing

Metrics and Validation

Audio Quality	Dropped packets (%) < 5%	Count original and final number of packets after transmitting an audio stream	Decrease audio packet size Choose another transport protocol
Battery Life	Hours of continuous use > 2hrs	Run device under heavy load for a set time to find battery usage	Add more battery Power with outlet

Metrics and Validation

Transcript Accuracy	Word Error Rate (%) < 25%	Check transcript for word error (substitution, deletion, and insertion) after speaking a known text*	Alternative models Switch to paid services (Google, Microsoft, IBM) NLP postprocessing
Speaker Identification Accuracy	Speaker Identification Error (%) < 25%	Check transcript for identification error after conducting a conversation with known contents and speaker switches	Alternative models Switch to paid services
Formatting Accuracy (chronology and speaker ID tags)	Formatting Error Rate (%) < 5%	Check transcript for formatting error instances after conducting a conversation with known contents and microphone switches	Augment metadata sent to meeting manager for merging transcripts

*100+ words. Well-formed sentences featuring common English words. Constant over multiple tests.

Project Management

	Phase 1 - Setup (weeks 4-6)	Phase 2 - Backend (weeks 5-8)	Phase 3 - Frontend (weeks 7-10)
Mitchell	Platform Setup	Audio Processing	Multi Speaker
Ellen	Machine Learning Integration	Meeting Setup, Speaker ID	
Cambrea	Hardware Setup	Audio Networking	Website Networking