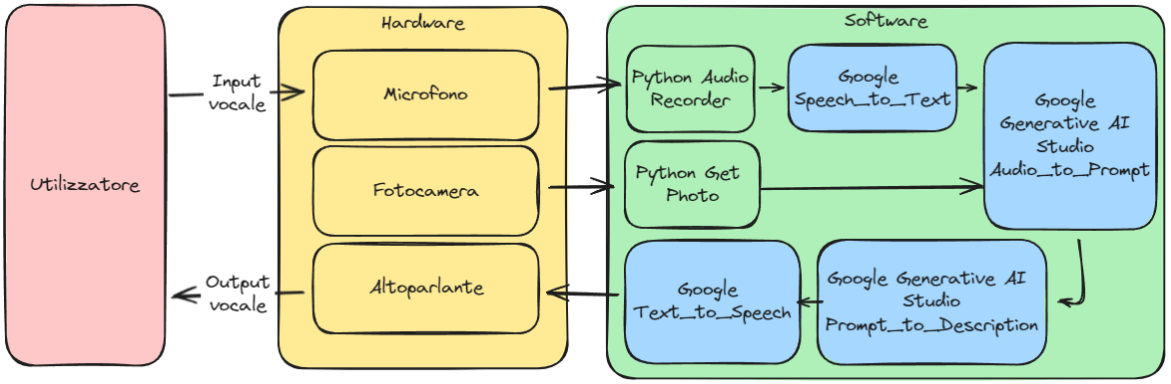
Architettura del dispositivo

L'architettura del sistema si basa su due componenti principali:

- Hardware: E' necessario che il dispositivo comprenda un microfono per l'acquisizione dell'input vocale dell'utente, una fotocamera per catturare immagini dell'ambiente e un altoparlante che permetta la riproduzione audio della risposta.
- Software: Sfruttando le potenzialità delle API Google di Intelligenza Artificiale Generativa, utilizza l'input vocale dell'utilizzatore e l'input visivo dell'ambiente per generare una risposta alla domanda dell'utente partendo dalle informazioni ottenute dall'ambiente circostante.

Funzionamento generale:

L'utente interagisce con il sistema tramite comandi vocali. Il microfono registra l'input vocale, che viene quindi elaborato dalla componente software. La fotocamera acquisisce immagini dell'ambiente in base ai comandi dell'utente. Il software elabora queste immagini e genera una descrizione vocale dettagliata dell'ambiente, che viene riprodotta tramite l'altoparlante del dispositivo.



Componente software

La componente software del sistema è responsabile delle seguenti funzioni:

Registrazione audio: Un modulo Python, denominato "Python Audio Recorder", si occupa della registrazione dell'input vocale dell'utente.
Acquisizione immagini: Un altro modulo Python, "Python Get Photo", gestisce l'acquisizione di immagini tramite la fotocamera predefinita del dispositivo.
Trascrizione vocale: Il servizio "Google Speech-to-Text" converte l'input vocale dell'utente in testo.

Generazione di prompt: Il testo trascritto viene elaborato per generare un prompt significativo per il modello di intelligenza artificiale Generativa che avrà il compito di generare la risposta a partire dal prompt e dall'immagine.
Descrizione dell'immagine: Il servizio "Google Generative AI Studio" utilizza il prompt generato per analizzare l'immagine acquisita e generare una descrizione dettagliata dell'ambiente.
Sintesi vocale: Il servizio "Google Text-to-Speech" converte la descrizione testuale dell'immagine in output vocale, che viene riprodotto tramite l'altoparlante del dispositivo.

Flusso di lavoro

Il sistema opera nel seguente modo:

1. L'utente fornisce un comando vocale, ad esempio "Aiutami a capire dove mi trovo".
2. Il modulo "Python Audio Recorder" registra il comando vocale.
3. Il servizio "Google Speech-to-Text" trascrive il comando vocale in testo.
4. Il software elabora il testo e genera un prompt per "Google Generative AI Studio", ad esempio, partendo da "Aiutami a capire dove mi trovo" ottengo "Genera una descrizione verbale dettagliata dell'ambiente circostante sulla base dell'immagine fornita. L'obiettivo è fornire informazioni utili e significative per aiutare i non vedenti a comprendere il contesto e le caratteristiche dell'ambiente. Assicurarsi di includere dettagli come la disposizione degli oggetti, la presenza di persone o animali, le caratteristiche dell'architettura circostante, il terreno e altre informazioni rilevanti per consentire una comprensione completa e sicura dello spazio. Includete anche l'avvertimento di alcuni potenziali pericoli. Iniziare la frase con ″Si trova....″. Non includere la descrizione dei suoni."
5. Il modulo "Python Get Photo" acquisisce un'immagine dell'ambiente.
6. "Google Generative AI Studio" analizza l'immagine e genera una descrizione dettagliata basata sul prompt.
7. Il servizio "Google Text-to-Speech" converte la descrizione testuale in output vocale.
8. L'output vocale viene riprodotto tramite l'altoparlante del dispositivo, fornendo all'utente informazioni dettagliate sull'ambiente circostante.