

音声によるWeb操作をエッジ処理で行う実現性検証

4DCS

武田 英士

takeda@office.4dcs.info

Web操作における不合理

Webサイト/アプリのUIというのは、詰まるところ、APIをラップして手続き化したものを、ユーザーに対して提供している。この中でユーザーが要求される行動は、あらかじめ用意されたシナリオから、目的につながるものを探すということ。

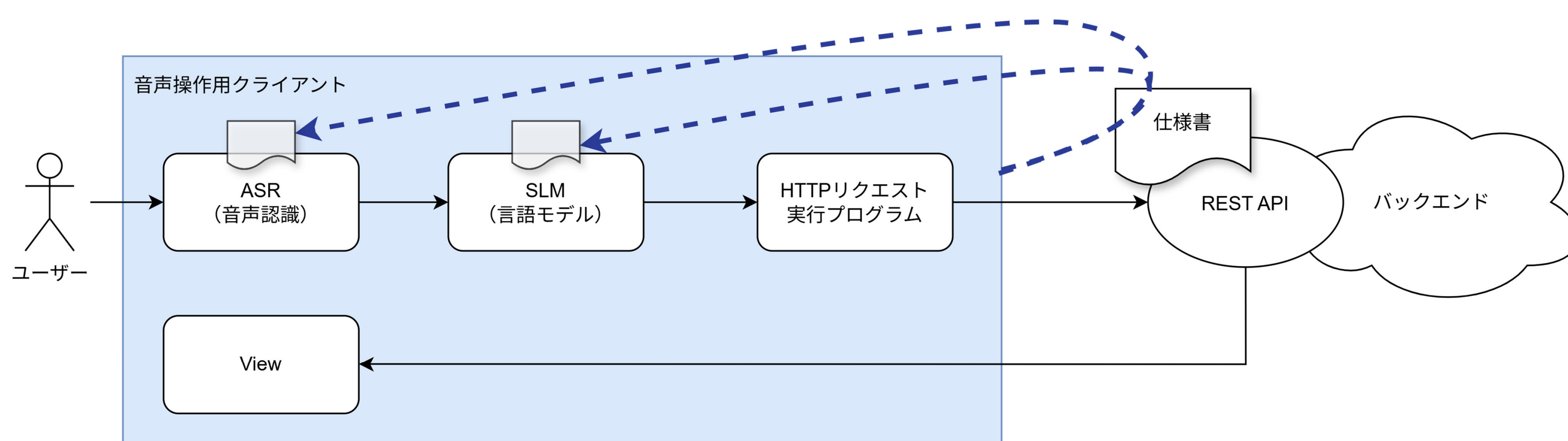
提供側もユーザーも目的に対して遠回りをしている。

音声認識とLLMによる解決

Web API(REST API)のインタフェースの定義仕様として、OASというものが標準化されている。これに則って記述された定義を事前知識を取り込んだ音声認識と言語モデルで、ユーザーの発話から、意図を解釈し定義に基づくリクエストに変換することができるのであれば、合理的なWebサイト/アプリにおけるインタラクションを実現できるのではないかと考えた。

実施した検証の概要

ブラウザの代わりに用いる、音声操作用クライアントを実装し、発話によって、CurlコマンドもしくはPostmanのようにWebアプリ(のバックエンドAPI)への直接リクエストを実行する。



なお以下のような要求の限定によって、

- 世界知識を必要とせず、ドメインが閉じている
- 語彙が限定される
- 意図パターンが有限

音声認識と言語モデルによる処理はエッジ側での実装が可能であろうと考えた。(起案時)

検証結果

検証において以下のような障壁が明らかになった。

- 仕様書から得られない情報の発覚
モデル知識に無い固有名詞において、標準仕様書からは“読み”の情報は得られない為、発話から定義された操作に解釈されない。
- 定義ファイルを拡張することで対応した
- 意図と一致しない操作の発生
操作がSLM任せなので、オペレーションやパラメータの誤選択・誤入力には常に発生しうる。
- 決定論的プログラムによる検査プロセスと確認インタラクションを実装し対応した
- ASRでの日本語対応の限界
日本語は、同じ音列に対応する漢字・表記の候補が多いため、音を認識するだけでなく、文脈から適切な文字表記を選ぶ必要があり、ASRでのボトルネックが、この構成においては何よりも大きかった。

結果評価と今後の展望

制作終盤はASRでのボトルネック解消のため、モデルサイズの変更も含め、チューニングに明け暮れたが、結局満足のいく精度は出ず、この問題の解決には至らなかった。
検証結果から、現時点では、音声によるWeb操作をエッジ処理のみで実現することは困難であると結論付ける。

起案当初の『エッジで実装可能』という部分がまず見込み違いだったので、この制約を捨てて、引き続き合理的なWeb操作の実現は考えていきたい。

また、ChatGPTやClaudeといった生成AIサービスの利用を前提に置くのであれば、MCPの活用など、生成AIの介在を前提とした、新しいWebデザインの在り方を模索していくような取り組みになると考えている。