スマートフォン向けにカスタマイズが可能なサイレントスピーチインタフェース
― 音声入力の自在化を目指して ―

1. 背景

音声入力では自然言語を媒介として人間とコンピュータが直接対話できるため、特殊な学習が訓練などを必要とせず、誰でも簡単に使うことができる。例えば、キーボードやタッチスクリーンなどの伝統的な入力方式を使う場合、ユーザはショートカットやボタンの配列を覚えないといけないが、音声入力ではその過程を必要としない。近年、深層学習に基づいた音声認識は革新的な進歩が起きており、Amazon 社の Alexa や Apple 社の Siri などの市販のボイススマートアシスタントが簡単に入手できるようになった。「アレクサ、寝室の電気をつけて」のような一言で、目の前の仕事に集中したまま並行して行える入力方式は人々の生活をより一層便利なものにしている。

しかし、音声入力の「発声を必要とする」という固有性質のために、環境の影響に弱い側面があり、3 つの課題が挙げられる。第一に、公共の場において、発声は周囲に迷惑をかけることや、個人情報の漏洩を引き起こす可能性がある。第二に、音声認識の精度がノイズに左右されてしまうため、特に騒音のある環境や距離が離れている場合に音声が認識されにくく、ユーザが大声を出さざるをえないこともある。第三に、ユーザ以外の人の声にも反応するなど、セキュリティ面のリスクが潜在する。

2. 目的

本プロジェクトの目的は、音声入力の自然さを保ったまま、プライベートなコミュニケーションができるサイレントスピーチによる入力方式を、スマートフォンなどの携帯端末で利用できるシステムとして開発することである。さらに、コマンド登録の手間を最小限に抑え、万人が使える認識システムとして実現することを目標とした。

3. 開発の内容

3.1. 対照学習を用いたリップリーディングモデルの作成

本プロジェクトでは、リップリーディングに基づいたサイレントスピーチの認識手法を開発した。正確に発話を認識できる深層学習モデルを作成するために、自己教師あり学習手法を用いて事前学習を行い、口元映像から効果的に特徴を抽出するエンコーダを実装した（図1）。事前学習は大規模のリップリーディングデータセット LRW 上で行うことで、様々な顔向きや照明環境、手ブレなどの影響にロバストなモデルを実現した。

3.2. ワンショット転移学習と iOS アプリの開発

特徴抽出器で抽出した唇の特徴ベクトルは発話の高次元の表現である。従って、シンプルな線形分類器を使うことで、少ないサンプルでも簡単に分類ができる。本プロジェクトでは、iOS の Create ML フレームワークの MLLogisticRegressionClassifier を用いて、特徴抽出器の出力（長さが 500 のベクトル）のノルムが 1 になるように正規化した上で口唇の動きの学習・予測を行った。

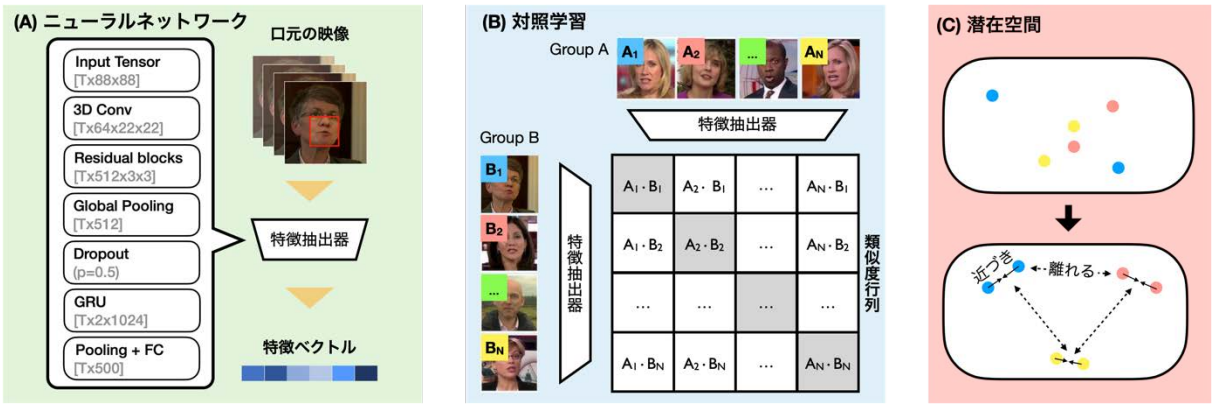

図 1:対照学習のパイプライン

スマートフォンに向けプライベートな入力方式を目指して、iPhone や iPad など iOS 端末で使えるリップリーディングシステムを実装した。図 2に示すように、カメラプロセスによって録画したビデオから口唇部を切り抜き、リップリーディングモデルで認識されたコマンドがショートカットとして実行される。また、コマンド登録の手間を最小限に抑えるために、音声認識によるコマンド登録を可能にするボイストゥーリップ(Voice2Lip)を開発した。使いたいコマンドを有声発話で 1 回話すことで、音声信号と口唇映像を同時に記録することができるため、音声認識の結果をラベル、口唇映像から抽出した特徴ベクトルを入力データとして、自動的にサイレントコマンドとして学習できる。その枠組みによって、数秒で新しいコマンドが使えるようにした。

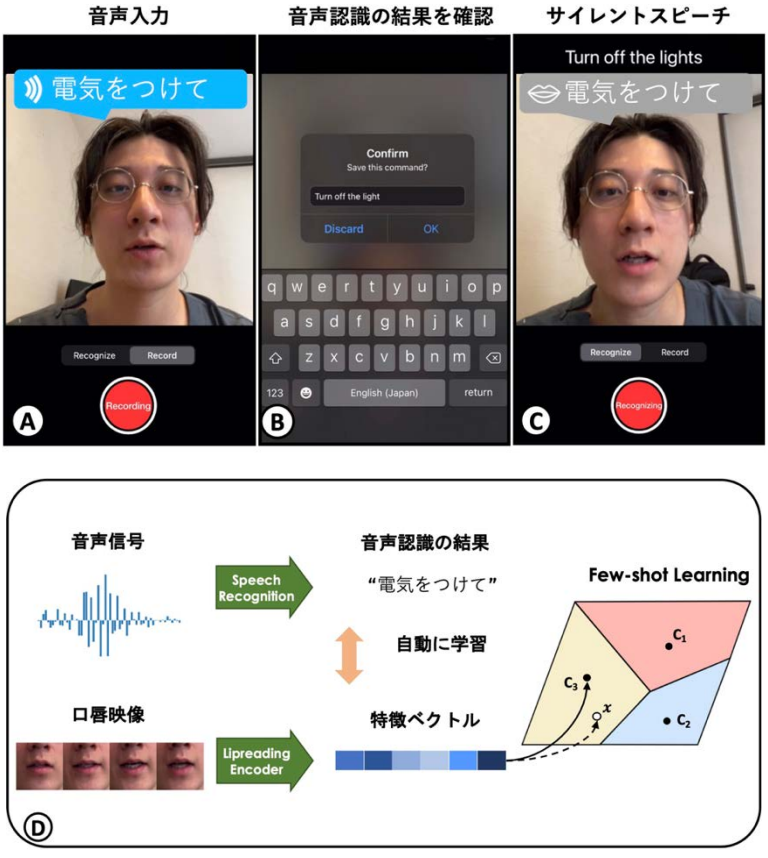

図 2:音声入力によるワンショット転移学習のパイプライン

3.3. キーワードスポッティング機能の開発

既存の研究やプロダクトでは、ほとんどがボタンを押すことで音声認識を開始させる仕組みになっているが、これだと手による操作が必要になり非常に不便である。図 3 に示すように、本プロジェクトでは類似度によるワンショットキーワードスポッティング（One-shot Keyword Spotting；One-shot KWS）機能を開発し、世界初のハンズフリーで無声発話認識を開始させることができるサイレント・ウェイクアップ（Silent wake-up）を実現した。

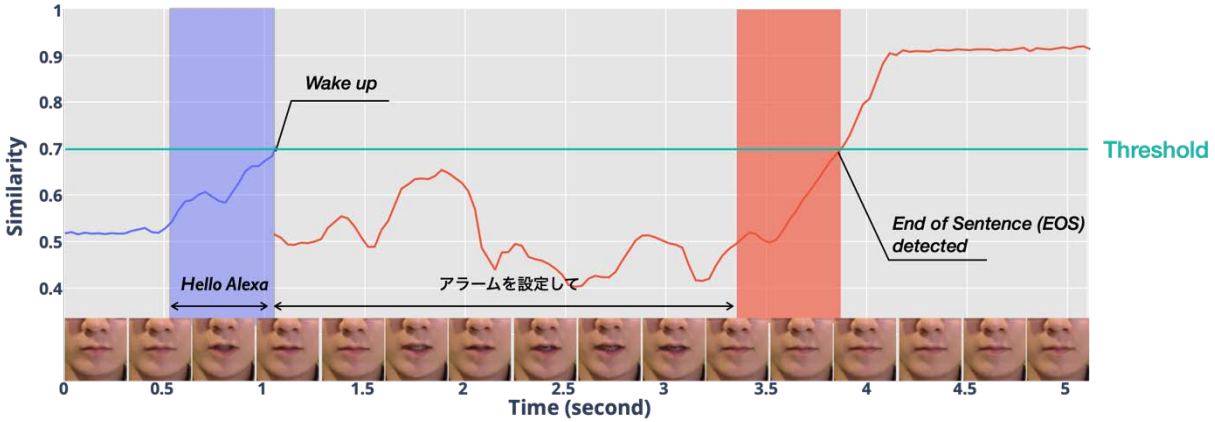

図 3：類似度によるワンショットキーワードスポッティング

3.4. ユーザテスト

実際に一般のユーザが使用する場合に、どれほどの精度で認識できるかを検証するために、ユーザテストを実施した。図 4（左）に示すように、KWS 機能の偽陽性率は 0.26%を達成し、ユーザが誤作動を報告するにつれて効果的に軽減されることを確認した。リップリーディングの結果を図 4（右）に示す。学習データが 1 サンプルのみの場合においても 81.67%という高い認識精度を達成した。さらに、使用するたびに自動的に新しいデータを学習することにより、3 ショットの精度が 96.04%、5 ショットの精度が 98.75%のように、次第に認識精度が上昇することも示された。また、多様な異なる言語に対しても性能の差異がないことも確認できた。

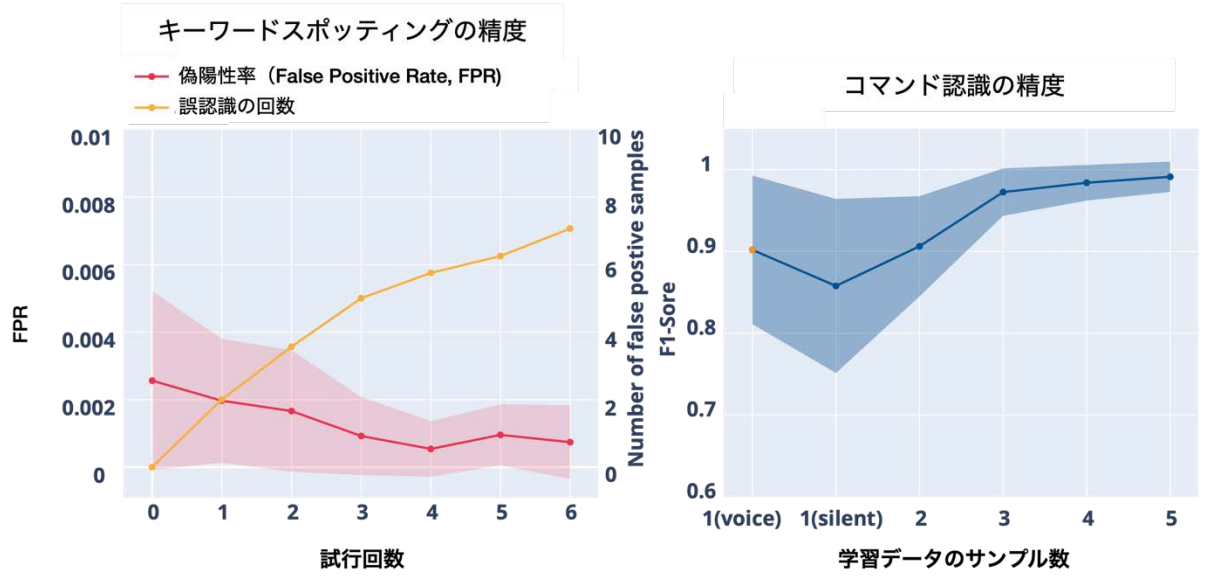

図 4：ユーザテストの結果

4. 従来の技術（または機能）との相違

既存の手法は利用可能なコマンド数が限られており、既定された単語やフレーズしか使えない。同時に使えるコマンド数が少ないため、サイレントスピーチだけでは一部の機能しか使用できない。すなわち、既存のリップリーディングインタフェースは発話によるインタラクションの最大の利点である「直感さ」と「自然さ」を喪失してしまっている。 また、新しいユーザが利用しようとするたびに学習データを収集する必要があるが、その所要時間は長くユーザに大きな負担をかけることに加えて、モデルの学習も一般的に数時間のトレーニングを要する。

本プロジェクトで開発されたリップリーディングシステムは、事前学習とワンショット学習によるファインチューニングを組み合わせることで、高い認識精度を保証しながら、1 つのサンプルで簡単に自分のコマンドを追加できる。本プロジェクトによって、リップリーディングの実用化と、誰でもすぐに利用を始められるサイレントスピーチを実現した。

5. 期待される効果

本プロジェクトにおいては、カスタマイズ可能なサイレントスピーチインタフェース実装し、音声入力の直感さを保ちつつ、公共の場においても自由に使えるサイレントスピーチインタラクションを可能にした。様々な場所に持ち込む携帯端末にリップリーディングシステムを実装することで、手や音声に依存しない斬新なモバイルインタラクションが実現できると予想される。

6. 普及（または活用）の見通し

本プロジェクトでは、サイレントスピーチ認識と iOS 上のショートカット機能を連結させ、実際にスマートフォンを操作できるアプリを開発できた。学習済みのモデルとソースコードをオープンソースで無償公開しており、多くの人にサイレントスピーチが利用できる環境を築いた。これからは、社会実装を通じて、産業における様々な実際の課題を解決し、より大きな価値を生み出すことを目指す。例えば、建設工事やインフラ整備などの作業現場では、機械制御や業務報告の利便性を向上させるために音声認識システムを導入しているが、機械の稼働音や路上ノイズなどで正確に音声を認識させることが困難である。リップリーディング技術を活用することで、手による作業を阻害せずに、騒音のある場所でも安定して音声を認識させられるシステムの開発を検討している。他にも、発話できない人のためのスピーチ入力システムなど、障害者支援の課題に取り組みたいと考えている。

7. クリエータ名（所属）

- 蘇 子雄（東京大学大学院）
- 方 詩涛（東京大学大学院）

（参考）関連 URL

ソースコードを公開している GitHub リポジトリ：https://github.com/rkmtlab/LipLearner