IPA

2022 年度 未踏 IT 人材発掘・育成事業 採択案件評価書

1. 担当 PM

藤井 彰人
（KDDI Digital Divergence Holdings 株式会社　代表取締役社長／KDDI 株式会社　執行役員　ソリューション事業本部　ソリューション事業企画本部）

2. クリエータ氏名

蘇 子雄（東京大学大学院学際情報学府 学際情報学専攻）
方 詩涛（東京大学大学院工学系研究科 電気系工学専攻）

3. 委託金支払額

2,736,000 円

4. テーマ名

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

5. 関連 Web サイト

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

6. テーマ概要

本プロジェクトでは音声不要の、誰でも自由自在に利用できる無声発話（サイレントスピーチ）インタフェースを開発した。具体的には、スマートフォンの内蔵カメラを用いた利用者の口元画像を元にリップリーディングを行い、発声を必要としないサイレントスピーチ入力を実現した。

従来のリップリーディングシステムはデータ収集に膨大な手間がかかったり、使用可能な語彙数も限られていたりするなどの課題が存在する。本プロジェクトでは、One-shot 転移学習を用いたリップリーディングモデルを実装し、大規模なデータセットで事前学習を行うことよって、1 サンプルだけでコマンドを登録できるリップリーディングシステムを実現した。これにより、語彙数の制限が解消され、サイレントスピーチコマンドをその場でカスタマイズすることが可能になった。

このリップリーディングによる認識手法とモバイル端末のボイスアシスタント機能を連動させることで、モバイル端末で気軽に利用できる、直感的で表現力の高い無声発話による入力を実現した。

7. 採択理由

音声インタフェースは、今やどこでも誰もが使えるインタフェースとして普及しているが、発話を前提とするため、騒音の影響を受けたり公共の場での発話が難しいなど、環境面での制約が多いのが課題である。

本提案は、リップリーディングに基づいたサイレントスピーチインタフェースをスマートフォンに実装することを目指している。具体的には One-Shot 転移学習を用いてリップリーディングを実装するとともに、スマートフォンのカメラとマイクを利用した個々にカスタマイズ可能なサイレントスピーチコマンド機能も計画しており、発話を前提とする音声インタフェースの「次」を担うことができるユニークな提案と考え採択した。

リップリーディングとスマートフォンで，これまでに経験したことない新しい世界を開いてくれることを期待したい。

8. 開発目標

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、サイレントスピーチによる入力方式を、プライベートな形態で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、万人が使える無声発話認識システムとして実現することを目標とした。

9. 進捗概要

本プロジェクトでの開発は、環境に対してロバストなリップリーディングモデルの作成と、One-Shot 転移学習を利用した iOS アプリの開発、そしてキーワードスポッティング等の UX 向上のための開発から構成される。単にリップリーディングモデルを開発するだけでなく、スマートフォン上で実用に耐えうるアプリケーションの開発を目指した。

対照学習を用いたリップリーディングモデルの作成では、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ている（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、様々な顔の向きや照明環境、手ブレなどの環境影響にロバストなモデルを実現した。

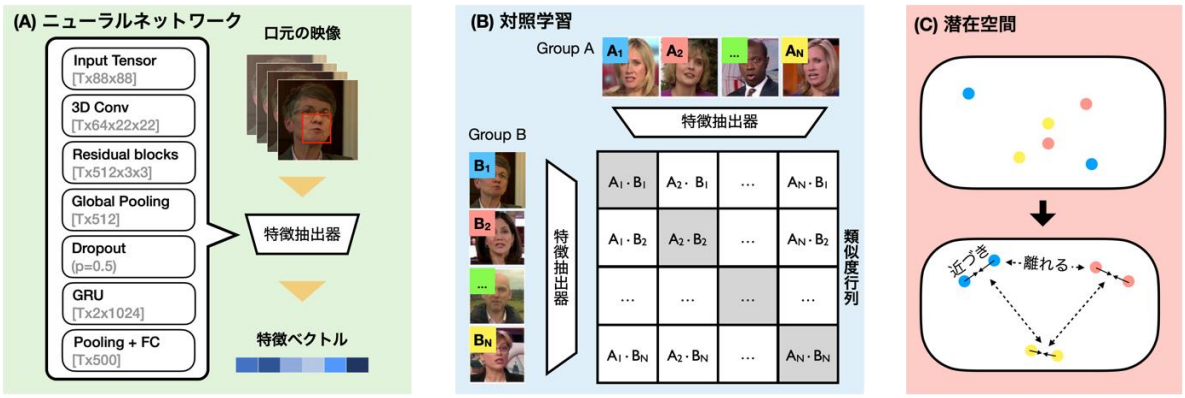

図 1：対照学習によるモデル作成

ワンショット転移学習と iOS アプリの開発では、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。

iOS 端末で、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アプリケーションを開発した。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 Voice2Lip を開発した。登録したいコマンドを有声発話で一度話せば、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として、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同時に利用することで、サイレントコマンドとして素早く登録することが可能となった。

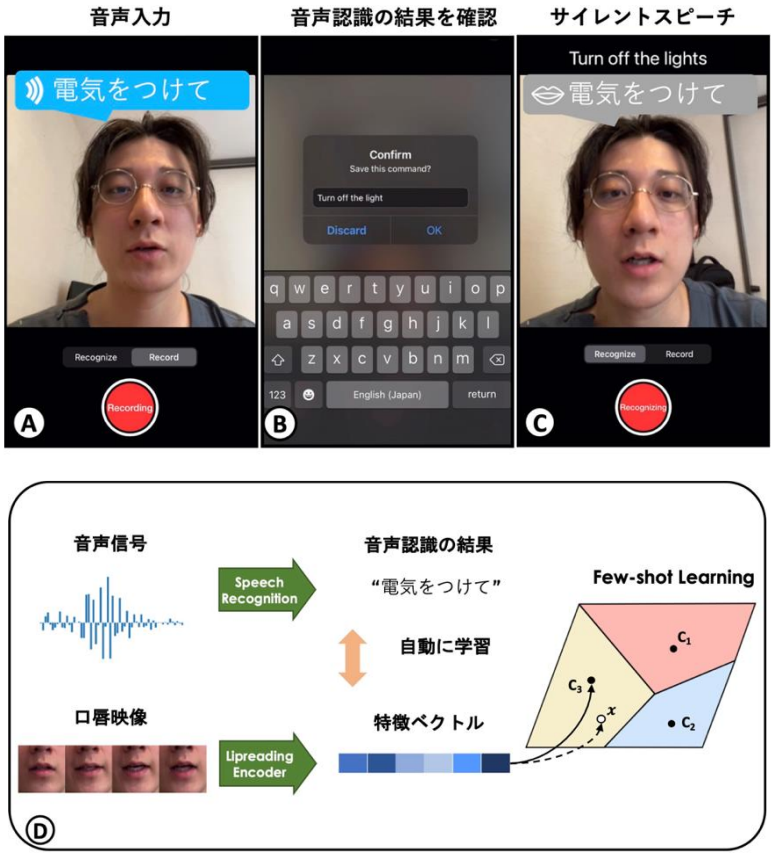

図 2：音声入力と転移学習による簡単コマンド登録

キーワードスポッティング機能は、サイレントスピーチ前の、ウェイクアップを可能する。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。

10. プロジェクト評価

スマートフォンやカメラがどこにでもある時代の、新たな入力システムを、本プロジェクトは見事に提示してくれている。リップリーディングですべての文言を完全に認識するという方向でなく、一回の発話で、数多くのコマンドをサイレントスピーチへ正確にマッピングするというのは、素晴らしいアプローチである。iPhone 上に実装したアプリケーションでは 30 個のコマンドを 98.75%の精度で認識することを実現しており、技術的な成果についても触れておきたい。iPhone/iOS 上での制約条件があるなか、これだけの精度を実現していることは高く評価されるべきと考えている。

中国出身のクリエータとして、はじめから日本語、中国語、英語の三カ国語に対応し開発を行ったこと、また、iPad でのデスクトップアプリ PowerPoint のショートカットをサイレントスピーチで動かすデモまで披露したことは、想定

以上の成果である。

蘇氏は日本語が流暢だとはいえネイティブではなく、また日本人クリエータもチームにいない状況において、本プロジェクトでは、日本語中心で行われる未踏プロジェクトでは様々な苦労もあったことだろうと想像する。両氏が、中国出身クリエータのみでも、未踏プロジェクトにおいて素晴らしい成果が出せることを示してくれたことに感謝したい。

11. 今後の課題

複数コマンドにサイレントスピーチを自由に結びつけるアプリケーションを、実際に利用できる形で開発したことは技術的に大変素晴らしい成果ではあるが、現実には、ユーザにカメラを起動させた状態で本アプリケーションを活用してもらうためには、さらなる検討が必要であろう。

デモで提示した PowerPoint のショートカット活用などは、解のひとつだと思われるが、多くの人に実際に活用されるためには、さらに何が必要なのか、どのような価値を誰に提供すべきなのか、プロダクトマネジメントやビジネス視点での検討が必要であり、さらなる発展を期待したい。