EarCommand: “Hearing” Your Silent Speech Commands In Ear

ECE 69500 IoT: Paper Review | February 09th, 2023 | **Presenter:** Zhengzhe Zhu | **Reviewer:** Yuan-Yao Lou

1 Introduction & Motivation

The paper addresses the issues of silent speech interfaces, which are a type of technology that allows users to control and communicate with electronic devices using their speech without making any audible sounds. Thus, the authors propose a new earphone-based silent speech interaction method, called EarCommand. The motivation is the limitations of the existing silent speech solutions such as background noise or interference, restricted hand involvement, and the complex setup caused by extra sensor devices.

2 Problem

01. The limitation of the existing silent speech solutions, include (a) noisy environments, (b) privacy concerns, (c) the efficacy and efficiency of the interactions, such as the systems requiring users to hold the smartphone in a relatively stable position and are not applicable in some special situations, and (d) required additional sensors.
02. The approach to detecting the subtle speaking-induced deformations of ear canals
03. Hard to distinguish various types of silent speech commands from subtle channel response feature variations

3 Solution

01. Solve the detection of continuous ear canal deformations (caused by articulator motions) by an in-ear acoustic sensing technique and proposing an end-to-end silent speech interaction system via a customized earphone
02. Address the challenge of command recognition by adjusting the channel response feature variations with dynamic reference inputs and proposing a CRNN-based framework for the word- and sentence-level silent speech commands

4 Experimental Results

The evaluation metric of the performance in an automatic speech recognition system is WER (Word Error Rate). The evaluation scenarios include **(a) wake-up/sleep activities** which achieved 95.9% accuracy, **(b) a comparison of deep learning models** to show the CRNN-based framework outperforms CNN and Transformer in terms of WER, **(c) word-level recognition performance:** MIN 5.6%; MAX 14.58%; AVG: 10.02%, **(d) sentence-level recognition performance:** MIN 8.9%; MAX 15.9%; AVG: 12.33%, **(e) robustness** by considering different noise environment, different motion behaviors and etc., and **(f) the overhead and latency** that one command can be performed in ~1.4s.

5 Pros and Cons

Pros: Not require additional sensors and the performance in terms of WER outperforms CNN and Transformer. The robustness can be achieved by up to 20% of WER in 9 realistic scenarios. The latency is only around 1.4s.
Cons: Hard to memorize the relationship of syllables among different words in a single sentence. Besides, more syllables in a single word a long word, and a long command both decrease the accuracy.

6 Questions

01. According to the evaluation of sentence-level performance, what kind of data augmentation and synthesis techniques do you think could be helpful to improve performance? (as the authors indicate the learning ability is limited based on the variable length of inputs)
02. The paper evaluates the overhead and latency by comparing the proposed approach with Amazon Echo. Is the response time truly 5 ~ 8 seconds? How this information is measured? Can we find any related research works?