Demo: Leveraging Earables for Unvoiced Command Recognition

Tanmay Srivastava
tsrivastava@cs.stonybrook.edu
Stony Brook University
New York, USA

Prerna Khanna
pkhanna@cs.stonybrook.edu
Stony Brook University
New York, USA

Shijia Pan
span24@ucmerced.edu
University of California, Merced
Merced, USA

Phuc Nguyen
vp.nguyen@uta.edu
University of Texas, Arlington
Arlington, USA

Shubham Jain
jain@cs.stonybrook.edu
Stony Brook University
New York, USA

ABSTRACT

We demonstrate an ear-worn technology that recognizes unvoiced human commands by tracking jaw motion. The ear-worn system is designed to achieve continual unvoiced command recognition for robust human-computer interaction (HCI) applications. First, the system reliably extracts the jaw motion signals buried under the noise caused by head motion, walking, and other motion artifacts to track single secondary voice articulator (i.e., word). Then, learning from linguistics and human speech anatomy, we design a novel algorithm that localizes the phonemes in the command, and reconstructs the word. We evaluate the proposed system in real-world experiments with 15 volunteers. Our preliminary results show that the proposed system obtains a word recognition accuracy of 95.6% in noise-free conditions and 93.2% and 91.6%, while head nodding and walking.

CCS CONCEPTS

• **Human-centered computing** → *Accessibility systems and tools*; **Accessibility**;

KEYWORDS

Unvoiced speech recognition, Wearable devices, IMU sensing, Earable

ACM Reference Format:
Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. Demo: Leveraging Earables for Unvoiced Command Recognition. In *The 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '22), June 25–July 1, 2022, Portland, OR, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3498361.3538665

1 INTRODUCTION

With the emergence of speech recognition techniques, voice assistants are now becoming increasingly common in our day-to-day devices [1]. Although voiced interactions with personal assistant devices are intuitive, they can be unreliable in noisy environments while compromising our privacy. Further, voice-based systems do not accommodate people with speech disorders [4] who find it difficult to produce sound, even if their jaw and lip movements are intact. Our work is motivated by the question: *"Can we recognize unvoiced speech to enable voice-like intuitive interactions?"*

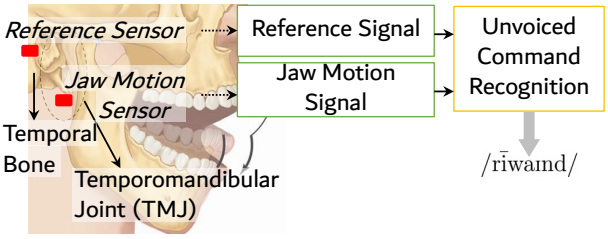

Figure 1: Concept and overview of our system

Unvoiced speech provides hands-free intuitive interaction. A recent study [5] shows that users prefer unvoiced speech over acoustic input as an interaction technique, and are willing to tolerate lower performance for maintaining their privacy. On the other hand, the increasing social acceptance of headphones and earphones leads to an alternative ubiquitous sensing modality – ear-worn sensors or *earables*.

In this demo, we present an earable system that recognizes unvoiced commands by tracking jaw motions. Compared to prior works, our system is capable of recognizing entire words/commands and is robust to body/head motions. Moreover, our system reconstructs a word from its components instead of training a word classifier, making the system scalable to a large commands dataset. We use a pair of inertial measurement units (IMU), one mounted behind the ear to be used as a reference sensor, and the other mounted on the TMJ to track the user's jaw motion. Figure 1 demonstrates our system's concept and overview. We demonstrate that these sensors can be integrated with a user's earphones to ensure a socially acceptable form factor.

We identify three research challenges for recognizing unvoiced command words using earables: (1) Indirect information inference. Speech production typically involves multiple articulators. It may be straightforward to interpret speech from the primary articulators

MobiSys '22, June 25–July 1, 2022, Portland, OR, USA

ACM ISBN 978-1-4503-9185-6/22/06…$15.00
https://doi.org/10.1145/3498361.3538665

since they are directly associated, however, secondary articulators are not distinctive enough for speech recognition. (2) Temporal phoneme overlap. As words are enunciated, phonemes tend to overlap to produce compound sounds. For example, in the word *mat*, the phonemes */m/* and */æ/* combine to produce the first part of the word, and the combination does not resemble either isolated phoneme. On the other hand, accurate word recognition would require precise phoneme isolation and identification, which is challenging when phonemes overlap; and (3) Body motion artifacts. Inertial sensors are susceptible to corruption by motion artifacts caused by body movements, such as head nodding or even walking. When tracking small jaw motion, large body movements lead to signal distortion.

We synthesize principles from linguistics with signal processing techniques to accurately detect unvoiced commands. Our system disintegrates the captured signals to phonological components, such as syllables, vowels, and visemes, which are composed of phonemes [2]. However, some phonemes are identical in terms of jaw motions. To overcome this, we model the task of word recognition as an estimation problem, wherein we reconstruct the word as a sequence of phonemes using a particle filter. The final output of the system is a list of phoneme sequences along with the posterior probability of each sequence. We validate the system for 55 words that are commonly used in voice commands. To ensure that only valid phoneme sequences are generated, we refer to a dictionary with 555 words.

2 SYSTEM OVERVIEW

Our system detects unvoiced commands by capturing jaw motions. There are 4 key steps. *First*, to isolate the jaw motions from head movements, we use twin-IMU sensing design (one on the temporal bone and the other on the TMJ). This allows our system to remove the head movement signal captured mutually by the twin-IMU. *Second*, we build a system that is not dependent on black-box machine learning algorithms to recognize words. It identifies the phonological components of the word called phonemes. The word is segmented first into *syllables*, and then the vowel within each syllable is localized based on the energy of the jaw signal. Next, our system identifies the first and last phoneme group of each syllable. *Third*, our system recognizes the unvoiced command with incomplete phoneme sequence using probabilistic modeling. We leverage a dictionary, which contains the phonemic map of 555 words (of these 55 are test words for which the system is evaluated), to ensure that only valid phoneme sequences are generated[1]. *Finally*, our system outputs a list of words (as phoneme sequences). This top-down approach enables us to reconstruct any word from its phonemic components.

3 DEMONSTRATION

Preliminary Results: We evaluate our system's overall command recognition accuracy and investigate its robustness in different environments. We invite 15 volunteers (4 females and 11 males) to collect data in an IRB-approved study. Our participants are between the ages of 19-31 years and speak different native languages — English (5), Hindi (4), Telugu (4), Spanish (1), and Kannada (1).

[1] Almost all standard dictionaries provide the phonemic map of each word [3].

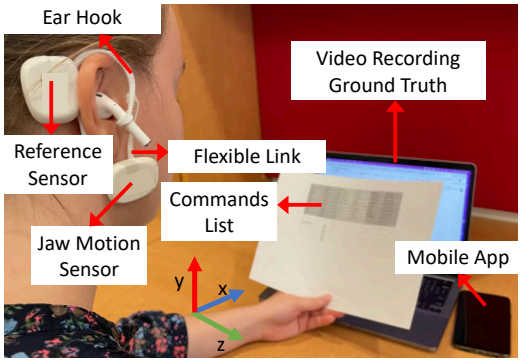

Figure 2: Our prototype and experiment setup.

We curate a set of 55 command words that are commonly used in voice assistants, smart homes, and gaming environments. From a randomized list of the command words, each user articulates the 55 words in an unvoiced manner 5 times and once in audible manner, with as minimal body movements as possible. We ask six users to collect additional data where they articulate the word while moving their heads and walking. Our system can achieve more than 95% word recognition accuracy in noise-free and more than 91% accuracy in noisy conditions (involving the user's body motion).

Demo: The goal of this demo is to show the usability and scope of an ear-worn device for unvoiced command recognition. One of the authors will wear the prototype as shown in Figure 2, and the system will recognize unvoiced commands. The end-to-end system will run on a battery-powered RaspberryPi. The twin-IMUs will stream data to the RaspberryPi device. We will invite the audience to pick the command, and the user will articulate the command in an unvoiced manner. The top 3 recognized commands along with their probability will be displayed on the screen.

4 ACKNOWLEDGMENTS

This material is based in part upon work supported by the National Science Foundation under Award Numbers 2110193 and 2132112.

REFERENCES

[1] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction* 26, 3 (April 2019), 17:1–17:28. https://doi.org/10.1145/3311956
[2] KENNETH De Jong. 2003. Temporal constraints and characterising syllable structuring. *Phonetic interpretation. Papers in laboratory phonology VI* (2003), 253–268.
[3] Collins Dictionary. 2021. Collins Dictionary. https://www.collinsdictionary.com/
[4] Madeline Jefferson. 2019. Usability of Automatic Speech Recognition Systems for Individuals with Speech Disorders: Past, Present, Future, and A Proposed Model. *undefined* (2019). https://www.semanticscholar.org/paper/Usability-of-Automatic-Speech-Recognition-Systems-A-Jefferson/73eefd141f43750b3ae0648e6ef099597e24c6c9
[5] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. *Acceptability of Speech and Silent Speech Input Methods in Private and Public.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445430