Leveraging Earables for Unvoiced Command Recognition

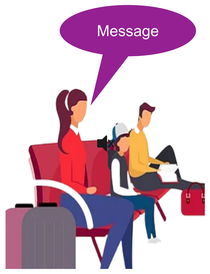
Tanmay Srivastava[↟] Prerna Khanna[↟] Shijia Pan[ʃ] Phuc Nguyen[+] Shubham Jain[↟]
[↟]Stony Brook University, [ʃ]University of California Merced, [+]University of Texas at Arlington

MOTIVATION

★ Voice assistants are limited by their unreliability in noisy environments and privacy concerns.

★ We present an earable to detect unvoiced words by capturing jaw movements using IMU.

Noisy Environment



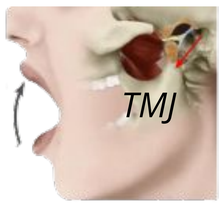
Private Interaction

HUMAN SPEECH ARTICULATION

★ Primary articulators
- Example: Lips, teeth, and tongue.
- Interact with other articulators to produce sound.

★ Secondary articulators
- Example: Jaw and nose.
- Cannot themselves make contact other articulators.

★ Jaw rotates about the TMJ to facilitate movement of tongue and lips.



KEY IDEAS

★ Recognize unvoiced speech from jaw motion.

★ Intuitive and unobtrusive input modality.

★ No ML-based word classification model.

★ Near-zero-effort scalability to recognize a large number of words.

★ Robust in presence of motion artifacts.

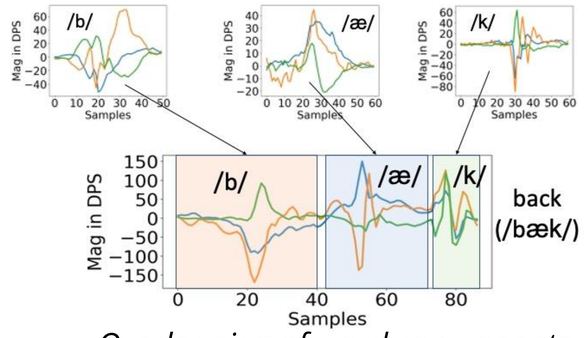
CHALLENGES

★ Detecting unvoiced speech from a secondary articulator.

★ Jaw motion is polluted by head and body motion.

★ Multiple sounds have similar jaw motion. Example: {*/m/, /b/, /p/*} and {*/t/, /k/*}.

★ Phonemes overlap to produce compound sounds.



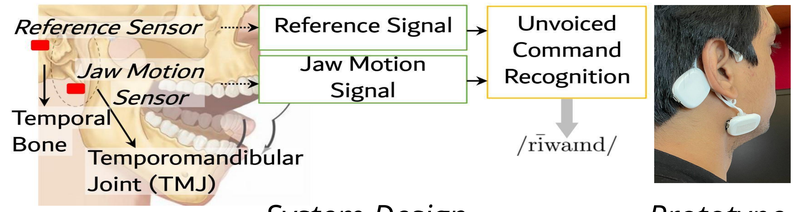
Overlapping of word components

SYSTEM OVERVIEW

★ Twin IMU setup to remove motion artifacts.

★ Disaggregate word signal into phonological components (syllables, vowels, visemes, and plosives).

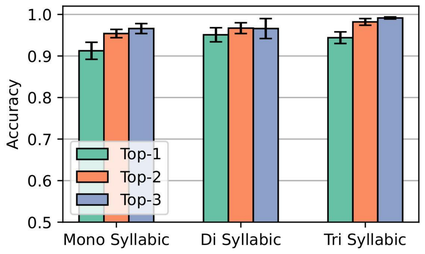
★ Reconstruct word from partial phoneme sequence using probabilistic estimation.



System Design

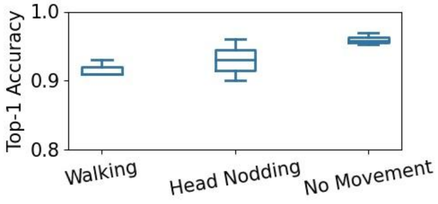
Prototype

RESULTS



Mean top-1 accuracy of 95.6% for 100 words across 15 users.

★ Accuracy when users are moving their head is 93.2%



★ Accuracy when users are walking is 91.6%