JawSense: Recognizing Unvoiced Sound using a Low cost Ear-worn System

Prerna Khanna [*#], **Tanmay Srivastava** [*#], Shijia Pan[‡], Shubham Jain[*], Phuc Nguyen[†]
[#]Authors with equal contribution, [*]Stony Brook University, [‡]UC Merced, [†]UT Arlington

Stony Brook University
Computer Science

OVERVIEW

JawSense is a wearable device that enables interaction with a computing machine based on unvoiced jaw movement tracking.

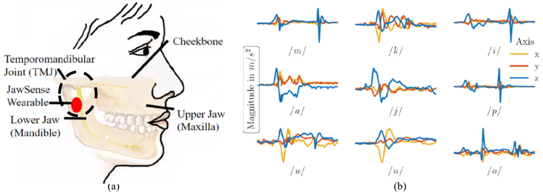

Fig. 1: ((a) JawSense concept and motivation. (b) Time domain representation of nine phonemes.

Speech and touch-typing are the two most common input modalities for smart devices. However, they pose challenges in many scenarios. We develop a low cost earable that captures jaw movements to decode 9 unvoiced phonemes spoken by a user.

CONTRIBUTIONS OF *JawSense:*

> Identifying ideal *sensor placement* for a single sensing modality to capture jaw movements.

> Proposing a new jaw sensing technique for hands-free, privacy-preserving, and *intuitive interaction* with devices.

> Algorithm to isolate unvoiced commands from voiced commands.

> Evaluating our system with six participants in a real world setting.

SYSTEM DESIGN

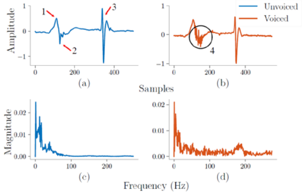

Fig. 2: a), (b) and (c), (d) are time and frequency representations of unvoiced and voiced phoneme /m/ respectively. It shows jaw opening (1), skin contraction (2), jaw closing (3), and voiced phoneme introducing high frequency component (4).

We place an accelerome-ter on TMJ as shown in Fig.1. It acts like a sliding hinge connecting the skull and the lower jaw. Placing the sensor lower on the TMJ, helps JawSense to retrofit on any off-the-shelf head-phone. We process the tri-axis accelerometer signal in 4 stages:

- Removing the effect of gravity and ripples.
- Removing the low frequency noise components like head movements, yawning etc. by high-passing signal with 1Hz.
- Mitigating the mechanical waves like external acoustic noise having frequency profile that overlap with the jaw movements. Assuming no more than one phoneme in two second window, more than one frame of half second with frequency range of [5,100] Hz is identified as event of acoustic noise. We use energy ratio to identify the frame with phoneme.
- We identify if the phoneme is voiced or unvoiced and chose not to detect voiced phoneme. When phoneme is voiced a high frequency, component is associated with jaw movement that originates from vocal cords while making sound that travel through cheek muscles. Using this property, we identify unvoiced and voiced phonemes as shown in Fig. 2.

EVALUATION

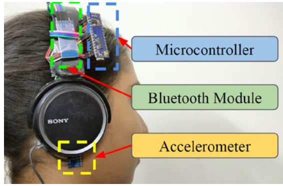

Fig. 3: JawSense Prototype

We develop a prototype as shown in Fig. 3, to evaluate the system on six subjects consisting of a microcon-troller, bluetooth module, and an IMU (MPU 9250).

We collect data at 550 Hz with 20 samples of each phoneme from each subject. We show that JawSense can be used as a form of hands-free and privacy-preserving human-computer interaction device with 92% phoneme classification rate in five folds cross validation as shown in Fig. 4, and 84% on leave one user out.

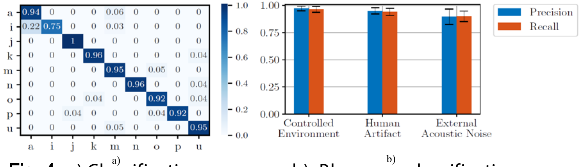

Fig. 4: a) Classification accuracy. b) Phoneme classification accuracy in different conditions.

CONCLUSION

We explored unvoiced sound recognition leveraging the motion of articulators involved in human speech generation. JawSense enables recognition of unvoiced phonemes via motion signals from jaw and cheek muscles.