Pantœnna: Mouth Pose Estimation for VR/AR Headsets Using Low-Profile Antenna and Impedance Characteristic Sensing

Daehwa Kim
Carnegie Mellon University
Pittsburgh, PA, USA
daehwak@cs.cmu.edu

Chris Harrison
Carnegie Mellon University
Pittsburgh, PA, USA
chris.harrison@cs.cmu.edu

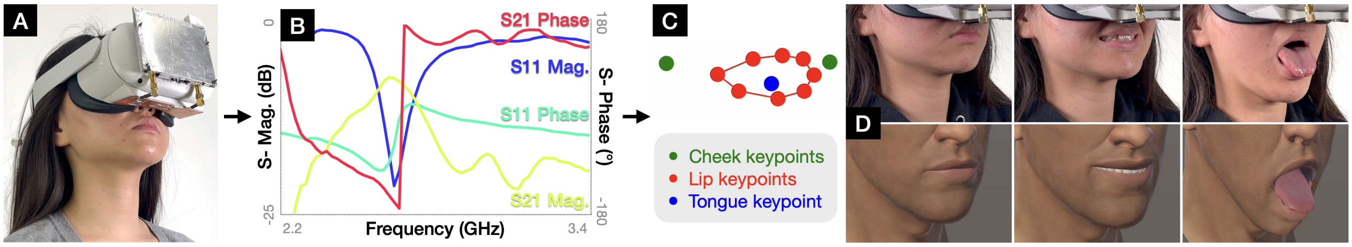

Figure 1: Pantœnna uses an antenna integrated on the bottom of a VR/AR headset (A). The mouth is dielectrically loaded to the antenna; changes in pose manifest as changes in the antenna's self-resonance frequency and performance, which we measure (B). Our machine-learning pipeline predicts 11 3D keypoints for the cheeks, lips and tongue (C). This data can then be used, for example, to pose an expressive avatar for telepresence uses (D). Our technique sidesteps privacy issues inherent in camera-based systems, while simultaneously supporting silent facial expressions that audio-based systems cannot detect.

ABSTRACT

Methods for faithfully capturing a user's holistic pose have immediate uses in AR/VR, ranging from multimodal input to expressive avatars. Although body-tracking has received the most attention, the mouth is also of particular importance, given that it is the channel for both speech and facial expression. In this work, we describe a new RF-based approach for capturing mouth pose using an antenna integrated into the underside of a VR/AR headset. Our approach side-steps privacy issues inherent in camera-based methods, while simultaneously supporting silent facial expressions that audio-based methods cannot. Further, compared to bio-sensing methods such as EMG and EIT, our method requires no contact with the wearer's body and can be fully self-contained in the headset, offering a high degree of physical robustness and user practicality. We detail our implementation along with results from two user studies, which show a mean 3D error of 2.6 mm for 11 mouth keypoints across worn sessions without re-calibration.

CCS CONCEPTS

• **Human-centered computing** → **Graphics input devices**.

KEYWORDS

Face, reconstruction, pose tracking, emotion, speech, avatars, social.

UIST '23, October 29-November 1, 2023, San Francisco, CA, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0132-0/23/10.
https://doi.org/10.1145/3586183.3606805

ACM Reference Format:
Daehwa Kim and Chris Harrison. 2023. Pantœnna: Mouth Pose Estimation for VR/AR Headsets Using Low-Profile Antenna and Impedance Characteristic Sensing. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29-November 1, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3586183.3606805

1 INTRODUCTION

Virtual and augmented reality (VR/AR) systems able to faithfully capture their wearer's holistic body pose will be key to enabling future telepresence and immersive social avatar experiences. A crucial component is the pose of the mouth, which people use to improve the recognition and bandwidth of spoken content [25, 49, 57, 60], as well as to glean emotive cues through facial expressions [18, 22]. When nonverbal signals such as facial expressions match situational context and spoken content, it can increase trust, clarity, and rapport among users [32]. However, if these cues are absent or mismatched, they can instead generate tension, mistrust, and confusion.

Considerable work has already been invested into mouth pose estimation, including worn computing systems such as VR/AR headsets. Indeed, consumer-oriented systems already exist with mouth tracking, most notably Meta's Quest 2 (audio-based) and Quest Pro (camera-based) headsets [51, 52], as well as VIVE's Facial Tracker accessory camera [68]. While already available on the market, each of these systems presents one or more downsides of note. Specifically, camera-driven methods are power- and computationally intensive, and more importantly, raise privacy concerns among some users. From Meta's public research [70], we can see that a headset-integrated camera captures the user's mouth and upper body, both of which are intimate areas that most users generally do not like to video up-close. Audio-based methods (e.g., [59, 67, 77] and what Meta uses in its Horizons app [53]) are perhaps less intrusive, but only work for speech and not for silent facial expressions. As we will discuss more in Related Work, researchers have looked into innumerable bio-sensing methods, but these generally require controlled contact with a wearer's skin and per-session calibration.

In this work, we describe a new approach for tracking the pose of a user's mouth when wearing a VR/AR headset. Uniquely among prior methods, it does not use camera or audio data, nor does it require any instrumentation of the user or sensors on the skin. Instead, our method uses a thin antenna that can be fully integrated inside the enclosure of a headset, allowing for compact form factors that are physically robust. Unlike audio-only methods, we can track mouth movements resulting from both speech and silent facial expressions. After reviewing key related work, we detail the implementation of a proof-of-concept prototype. We used both RF simulation and real-world experiments to hone our design. To validate our technique, we ran two user studies: one investigating the tracking of facial expressions, and the other focusing on spoken content. The main contributions of Pantœnna are:

- Exploration on a new application area – mouth pose – for impedance characteristic sensing.
- Successful demonstration of a low-profile antenna design for mouth pose sensing, allowing for much thinner integrations than prior work.
- Identifying new modes of operation that were not previously explored in related HCI work, including the use of dual-mode antennas, polarization, S21 data, and sensing directivity.

Our many experiments and results not only validate our current work, but also help inform future work, and in general, raise the technical sophistication of this technique in HCI.

2 RELATED WORK

In terms of applications of sensing mouth pose, the main uses have been in telepresence [21], facial motion capture for film and games [54], hands-free input [2], and diet monitoring [7, 26, 64]. Our present work does not contribute to new applications of mouth pose, but rather contributes a new technical approach. As such, we primarily review other technical methods for mouth pose, broken down into discrete classification vs. continuous tracking. Importantly, we only discuss on-body systems, and do not include methods using external infrastructure (such as fixed camera setups). Additionally, we do not review approaches for tracking other parts of the face, such as eye gaze direction [62], eyelids [30, 31], eyebrows [13], nose and ear movements. We conclude this section with a more focused discussion on prior systems that utilize the same sensing method. Table 1 provides an overview of key systems.

2.1 On-Body Systems Offering Discrete Mouth Pose Sensing

Most prior work on mouth pose digitization has focused on discrete classification of a limited number of key facial poses, such as smile or open mouth. Innumerable technical methods have been explored, and so we review this literature according to sensing modality.

Starting first with optical methods, researchers have augmented glasses [74] and VR headsets [1] with cameras for mouth pose recognition. In a more privacy-preserving fashion, Masai et al. [44, 45] and Suzuki et al. [66] used photo-reflective sensors to measure skin deformation caused by facial expressions.

Moving to mechanical vibrations and movements, CanalSense [2] used barometers in the ear canal to detect changes in air pressure resulting from face-related movements. IMUs can also be used – EarBit [7] proposed an ear-bud device with IMUs placed behind the neck and ear, along with a microphone and proximity sensor to detect deformation of the ear canal from jaw motion. Similarly, MyDJ [64] captured chewing signals using a piezoelectric sensor and an accelerometer. FitByte [6] embedded five IMUs into the frames of glasses (as well as a proximity sensor and a camera) to track food and liquid intake. Acoustic methods are also possible, such as that demonstrated in Interferi [29].

Electromyography (EMG), measuring muscle activations, is another popular way to detect facial gestures [8, 27, 28, 43, 48]. As one example system, Cha and Im [8] attached EMG electrodes to the face where a VR headset would rest and classified approximately nine mouth-related gestures. Similarly, Gruebler and Suzuki [27, 28] placed electrodes on the side of the face to detect smiles and frowns, in addition to a neutral class.

There is also a separate and large literature on tongue-based sensing. TongueSee [75] recognizes six tongue gestures using EMG electrodes on the skin. Saponas et al. [63] developed a dental-retainer-like device that recognized tongue-swiping gestures. Capturing

System	**Sensing Method**	**Discrete Pose** [# Classes, % Error]	**Continuous Pose** [Within-; Across-Session Error]
MeCap [1]	Camera	✓[5, 95.6%]	
Emoglass [74]	Camera	✓[7, 73.0%]	
CapGlasses [46]	Capacitive	✓[5, 89.6%]	
EarfieldSensing [47]	Electric field	✓[5, 90%]	
EarBit [7]	Inertial	✓[2, 93%]	
FitByte [6]	Inertial	✓[2, 83.1%]	
MyDJ [64]	Piezo + IMU	✓[2, 98.4%]	
CanalSense [2]	Barometer	✓[4, 87.5%]	
Masai et al. [44, 45]	Photo Reflective	✓[8, 78.1%]	
Cha et al. [8]	EMG	✓[9, 85.0%]	
Gruebler/Suzuki [27, 28]	EMG	✓[3, 85.0%]	✓(smile intensity)
Interferi [29]	Acoustic	✓[5, 68.4%]	✓(mouth open size)
VIVE Facial Tracker [68]	Camera		✓
Chai et al. [9]	Camera		✓
Wei et al. [70]	Camera		✓
NeckFace [10]	Camera		✓
C-Face [11]	Camera	✓[8, 88.6%]	✓[1.4 mm; 2.8 mm]
Li et al. [39]	Strain + Camera		✓
Luo et al. [43]	EMG	✓[5, 86.3%]	✓
BioFace-3D [72]	EMG + EOG		✓[2.4 mm; -]
Richard et al. [59]	Speech audio		✓
EarIO [40]	Acoustic		✓
Pantœnna (this work)	Antenna imp.	✓[10, 91.1%]	✓[1.8 mm; 2.6 mm]

Table 1: Overview of prior systems that are both worn and sense the external pose of the mouth. When possible, we include the number of recognized classes for discrete pose systems (sometimes including facial expressions that involve the mouth) and tracking accuracy for continuous pose systems. In some cases, we had to infer results and estimate values from figures. We further caution there are many system and evaluation specifics (different keypoints, pose sets, train/test details, etc.) that make direct comparison challenging. Please refer to individual papers for important details.

tongue movement has also been exploited for silent speech interfaces, such as SilentSpeller [36], which used electropalatography (EPG) for text entry. Using seven magnets attached to the tongue, lips, and teeth, and two magnetometers on the cheek, Fagan et al. [24] enabled recognition of 13 phonemes and 9 words. The multimodal sensing approach in Sahni et al. [61] used one magnet on the tongue combined with optical ear canal deformation.

Finally, and most similar to Pantœnna, are electrical mouth pose sensing methods. EarfieldSensing [47] uses electrodes placed inside the ear canal to detect both electric field changes and physical deformations resulting from mouth movements. Even more similar to Pantœnna is work by Rantanen et al. [58], which used non-contact capacitive sensors to measure proximity to the skin. The same method is used in CapGlasses [46], which achieves a fairly unobtrusive glasses-like design to which we also aspire.

2.2 On-Body Systems Offering Continuous Mouth Pose Sensing

Considerably more challenging than discrete classification of mouth state is continuous tracking of the mouth shape (i.e., mouth pose). Fewer approaches have been successful in this endeavor, which we now review. Please also refer to [71] for a recent survey of techniques, as well as Table 1 for an overview of highly relevant systems.

A common and highly-successful approach is to use one or more worn cameras operating in front of the face [39, 52, 56, 70]. Such systems are so accurate that we use VIVE's Facial Tracking camera accessory [68] as ground truth in our user studies. More advanced is Li et al. [39], which employed an RGB-D camera combined with strain gauges to better predict facial movements occluded by a headset. NeckFace [10] predicts 3D keypoints for eight facial expressions using a neck-worn camera. C-Face [11] uses two cameras located on the sides of the head to capture face contour deformations that occur during different facial expressions. As noted in our introduction, while these camera-based methods are highly capable, they also tend to be fairly power-hungry and raise privacy concerns with users.

Moving beyond cameras, another successful approach has been to use phonetic sounds in a user's speech to infer mouth pose [23, 59, 67, 77]. There are now commercial systems utilizing this approach, most notably the LipSync SDK [51] used in Meta's Quest product lineup. A significant drawback to this approach, however, is that silent facial expressions cannot be detected.

There are also a handful of mouth pose systems utilizing biosensing means. EMG is perhaps the most common — for example, Luo et al. [43] and Mavridou et al. [48] both demonstrated a VR headset with EMG electrodes located around the face pad for facial tracking. BioFace-3D [72] combines EMG and EOG (electrooculography) sensors into a wrap-around head sensor design. Interferi [29], also mentioned in the section above, used acoustic interferometry to estimate the continuous intensity of a smile. While promising, no other facial poses were investigated. Lastly, EarIO [40] emits audio from headphones towards users' checks, capturing distinctive reflections from changing face geometry to enable continuous pose tracking. Importantly, all of the above methods require consistent contact with the user and generally do not work well across worn sessions without recalibration.

In summary, all mouth pose methods suffer from one or more downsides, which makes this research problem very much an open question. To this literature, Pantœnna contributes a new sensing approach with a unique set of strengths and weaknesses, as we will discuss. It may also be that a multimodal approach will be key to making progress in this domain, and the compact nature of Pantœnna makes it a promising technique for future work.

2.3 Antenna Impedance Characteristic Sensing

Our approach is built around antenna impedance characteristic sensing, a technique that has been explored in prior work. First, Xu et al. [73] enabled four-finger gesture classification using an antenna located on the wrist. Li et al. [41] used a body-worn monopole antenna to classify human activities. In the HCI domain, AtaTouch [34] explored a V-shaped antenna embedded in a VR controller to precisely segment finger pinch events. More recently, this approach was used in EtherPose [33] for sensing continuous hand pose. Our work is directly inspired by EtherPose, and we both advance the technique and apply it to a new domain.

To summarize the overall approach, dielectrics (e.g., human tissue) close to the antenna are loaded as a component of the antenna and affect its reference ground plane, and consequently affect the antenna's characteristic impedance and radiation performance at a given reference frequency. Geometry changes in the dielectric (e.g., mouth movement) thus vary antenna performance and can be measured as S-parameters. In comparison to EtherPose [33], we make several advances and contributions. First, we designed and validated a new and more advanced low-profile antenna topology. Whereas EtherPose's used two volumetrically-large cloverleaf antennas, Pantœnna uses a sub-millimeter low-profile cross-polarized antenna system, which is far more practical for integration into consumer electronics. Second, we newly utilize S21 data and quantify the performance gain over using S11 signal alone. Third, we utilize better sensing hardware, unlocking higher framerates and resolution. Fourth, we explore a new application area for antenna impedance characteristic sensing — mouth pose — underscoring the potential generality of the technique. Indeed, much like capacitive sensing, we believe there is not just one application, but rather a significant range of potential uses waiting to be explored by the HCI community.

3 ANTENNA DESIGN

Before we could create a prototype system, we needed to better understand several important design parameters of our antenna. We now briefly describe our iterative design process.

3.1 Test Setup

We primarily used electromagnetic simulation software (CST Microwave Studio [17]) to make antenna system topology design decisions. To test the S-parameter response for each mouth pose, we built five representative mouth poses using a 3D human phantom head: *closed mouth*, *open mouth*, *smile*, *smile with teeth*, and *tongue out open* (Figure 2). We used commercial 3D EM phantoms for the head, lips, teeth, gum, and tongue. The VR headset mirrored

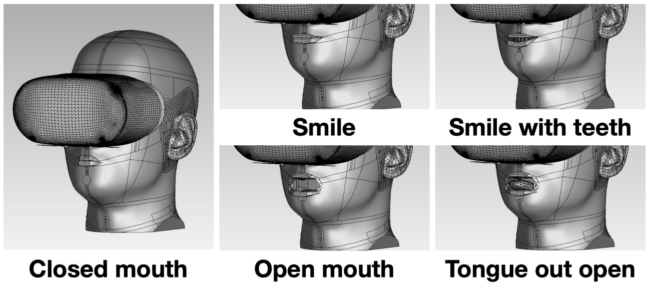

Figure 2: An EM phantom head wearing a VR headset expressing five representative mouth poses. The antenna is located beneath the headset. See also Figure 7 for illustrative photos of these mouth poses.

the dimensions of an Oculus Quest 2 [50], though the electronics inside were not modeled. Each part was assigned material properties found in SPEAG [65], which emulates the response to electromagnetic radiation. Lossy copper material properties were used for all antennas. The simulated antenna was located on the bottom side of the headset (seen in Figure 5). All of our simulations were run in a range from 500 MHz to 3 GHz.

To minimize the mismatch loss, we implemented an impedance matching network for each antenna at a predefined frequency of interest, where the magnitude of S11 is around -30 dB. Despite our efforts to match the simulation to the real world, there are limitations. For example, we did not model the inside electronics of the headset. Simplified human tissues and mouth shape on the phantom are also not identical to actual humans. Nonetheless, there were strong correlations between real-world and simulated results that allowed us to leverage computer simulations to expedite our iterative design process.

3.2 Antenna Topology

Our antenna is designed to minimize interference from the hands and environment with a directional radiation pattern. While there are many directional designs, we selected the slot antenna for its low-profile and simple structure (Figures 3, 4, and 5). To verify its performance, we simulated the signal response on five mouth poses. The design and its specific dimensions can be seen in Figure 3. The feed line is located 10 mm from the end of the slot.

Figure 3 shows the antenna impedance characteristic for four mouth poses. For illustration only, we use the *smile with teeth* mouth pose as a baseline and subtract all other mouth pose signals from this signal. We probed two center frequencies, 1.56 and 2.80 GHz using different impedance matching networks. For both frequencies, we were able to verify that both S11 magnitude and phase were responding to mouth geometry changes, yet the amount of signal difference varied. At 2.8 GHz, the signal change for the *tongue out open* gesture is very strong (~6.6 dB, ~30°), while the change in other mouth poses was around 1.2 dB. Conversely, at 1.56 GHz, all mouth poses exhibited significant and distinctive changes (~4.5 dB, ~30°), except for the *smile* pose (0.8 dB and 9°). Regardless, the antenna design offered excellent signals at both frequencies and so we moved on to investigating other parameters, leaving a more thorough investigation of operating frequency until later.

3.3 Two-Port Configuration and Polarization

Next, we considered a two-antenna design, which not only allows two antennas to measure their reflected signal (S11 and S22), but also measure how much signal is transmitted from one antenna to the other (S21 and S12). To accommodate two antennas, we folded a slot antenna in half so that we could include a second antenna on the same copper ground plane. Slot antenna polarization is determined by the orientation of the electric field across the slot, and thus we could also vary polarization of the two antennas. We tested two obvious layouts, illustrated in Figure 4, both with a center frequency of 2.8 GHz. In the co-polarized configuration, the antennas are the same shape and orientation, placed next to each other. In the cross-polarized configuration, the antenna at port 2 and its feed line are rotated at 90°, but the location of the feed line is consistent with the co-polar configuration (Figure 4, bottom).

Overall, we observed that the cross-polarized design was significantly more reactive to mouth pose changes, especially in the S21/S12 signal. Also, it is not surprising that having two antenna polarizations allows each antenna to be receptive to different mouth pose features (as opposed to having two identical, symmetrically-placed antennas). In general, co-polarized antennas have lower isolation, and therefore are less sensitive to changes of the ground plane due to changes in mouth shape. Conversely, better-isolated antennas (e.g., cross-polarized) have larger isolation dynamic range, and thus are more responsive to different mouth poses. For these reasons, we selected the cross-polarized design.

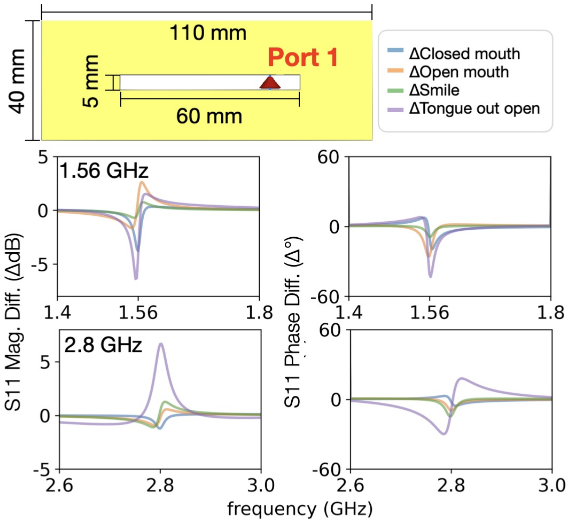

Figure 3: Simulation results of an antenna with a single straight slot (top, in yellow). Two center frequencies were tested, using different impedance matching networks. Each colored plot is the signal difference between a mouth pose and *smile with teeth* (see Figure 2).

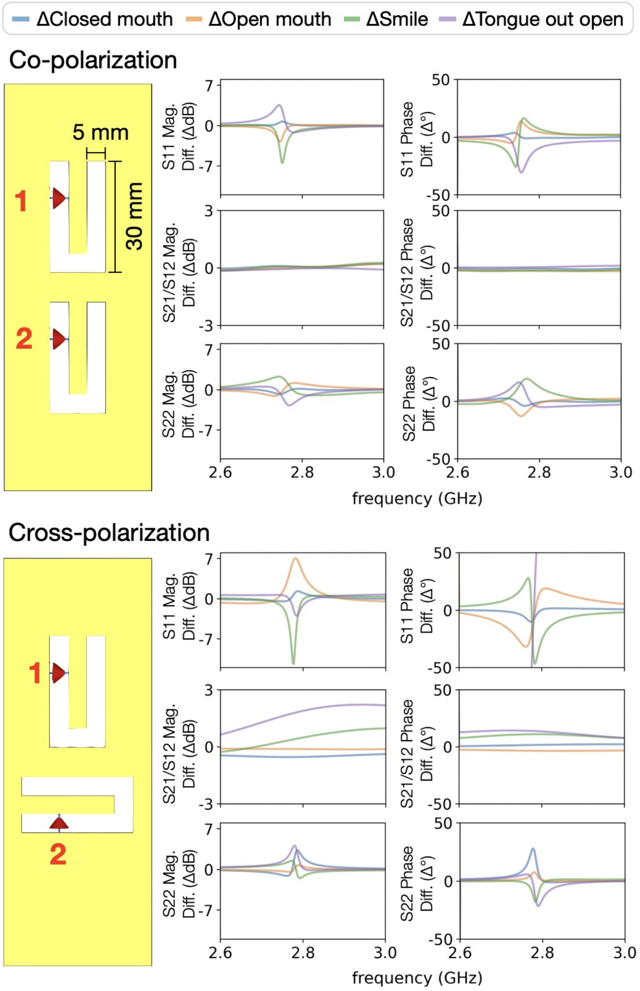

Figure 4: Antennas with two different polarization configurations were simulated (left, in yellow). Exterior dimensions were 40×110 mm. Each configuration includes two ports, and thus three signal combinations (S11, S21/S12, S22) are reported. Each colored plot is the signal difference between a mouth pose and *smile with teeth* (see Figure 2).

3.4 Operating Frequency

As a final exploration, we investigated the effect of operating frequency. For this, we used a variant of the cross-polarized antenna identified in the prior section. Specifically, we placed the antenna at port 1 in the center, with the second antenna shifted to the right (Figure 6). We call this design a Dual Asymmetric Enhanced Half-Wavelength Antenna. We found this arrangement could better capture asymmetric mouth movements, confirmed in both simulated and real-world experiments. We also angled the antenna slightly, using a 3D-printed wedge, to improve its directionality towards the mouth.

We considered four center frequencies of interest: 1.5, 1.97, 2.5, and 2.8 GHz. To tune the operating frequency, we used a different impedance matching network. In the simulation, we injected signal at port 1 (and measured S11), and port 2 was used only for measuring (S21). Overall, the simulation results show better discrimination at higher frequencies (Figure 5). At 1.5 GHz, there is very little signal variation across mouth poses. S21 starts to yield information at around 1.97 GHz, but S11 is largely unchanged. Both 2.5 and 2.8 GHz offer useful signals in both S11 and S21, with 2.5 GHz appearing to be a sweet spot, and thus we selected this as our operating frequency moving forward.

4 PANTŒNNA SYSTEM

The experiments in the previous sections informed the design of a proof-of-concept headset, seen in Figure 6. We now describe the major hardware and software components of our prototype, which we use in our subsequent user study.

4.1 Headset

As a proof-of-concept platform, we selected the popular Meta Quest 2 [50]. Although this headset can operate in an untethered fashion, we tether it to an M1 MacBook Pro 16" (2021) to simplify development and evaluation. In Future Work, we discuss avenues for commercial integration.

4.2 Antenna

We affixed our Dual Asymmetric Enhanced Half-Wavelength Antenna design to the bottom of the headset using a low-profile 3D-printed wedge (increasing from 0 to 8.5 mm in height), angled slightly towards the face. This design is cross-polarized – the vertically-polarized U-shaped slot antenna is located at the center, while the second horizontally-polarized antenna is placed on the right, 10 mm from the edge of the other antenna (see Figure 6). The antenna slots are 60 mm in length and 5 mm in width. To reduce antenna footprint, each slot is physically folded in half-length, maintaining the same electrical length as a straightened slot.

To fabricate this antenna, we first laser-cut our design out of a 1 mm acrylic sheet. We then overlaid this with copper tape and removed the voids with an exacto-knife. An SMA (Sub-Miniature Version A) coaxial cable was soldered to each antenna. The outer-conductor mesh of the coaxial cable is soldered to a ground plane formed by the copper tape, while the feed line is soldered across the antenna slot.

The self-resonance of this antenna design is 2.1 GHz with -23 dB of S11 magnitude. To shift the antenna's operating frequency closer to our target frequency of 2.5 GHz (identified in Section 3.4), we used an impedance matching network (designed in simulation; shown in Figure 6). Specifically, we connect a parallel inductor (30 nH) between the antenna feed line and ground, and a capacitor (0.6 pF) is serially connected to the port and feed line. While the simulation result of this matching network yields -27 dB of S11 magnitude at 2.52 GHz self-resonance, the actual antenna prototype has a self-resonance of 2.64 GHz and -22 dB of S11 magnitude.

4.3 Analog Front End

For measuring S11 and S21 parameters, we use a $230 NanoVNA V2 Plus4 [55], which is attached to the front of our headset prototype (Figure 6). As a brief primer, the S11 parameter is power reflected back to a transmitting antenna, while S21 is the transmitted power

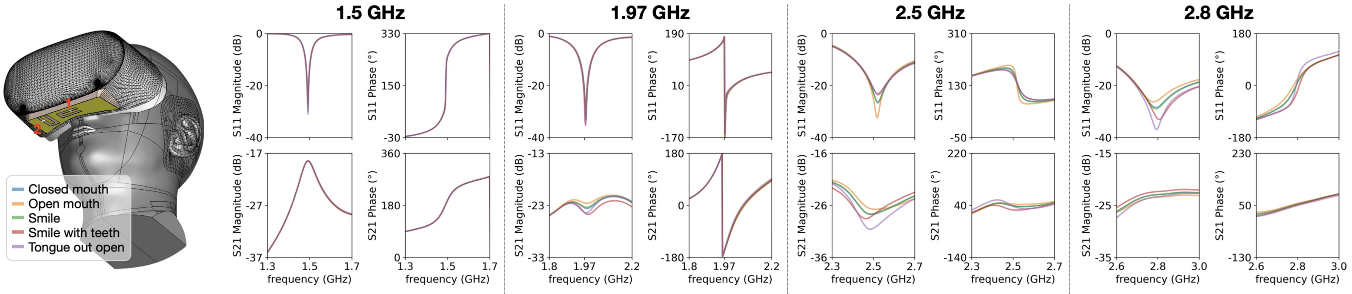

Figure 5: Four operating frequencies were tested using our final antenna design (placed on EM phantom head, far left; photo of the antenna in Figure 6).

between a first antenna and a second antenna. We configure our VNA to measure S11 and S21 from 2.2 GHz to 3.4 GHz, calculating return loss magnitude and phase shift. This particular VNA model is not able to measure S22, though other VNAs can and the signal is likely to be valuable. Nonetheless, the combined information from S11 and S21 is still sufficient to estimate mouth poses. We note this "nano" VNA is still a sophisticated measurement device. In a commercial implementation, a more basic, inexpensive, and specialized single-chipset integrated VNA design could be employed, and integrated onto the motherboard of the headset.

In piloting, we found that facial expressions did not require high framerates, but did benefit from a higher-resolution frequency sweep to better distinguish among diverse mouth poses. In response, we sample 61 points from 2.2 to 3.4 GHz (i.e., 20 MHz increments) at 5 FPS. This yields 61 S11 return loss magnitudes, 61 S11 phase shifts, 61 S21 return loss magnitudes, and 61 S11 phase shifts, for a total of 244 values, which we pass to our machine learning pipeline. Conversely, mouth movements during speech are more rapid, but require less resolution to capture the gestalt of the mouth moving. In response, we sample only 31 points from 2.2 and 3.4 GHz (i.e., 40 MHz increments), yielding 124 values at a faster 8.5 FPS. We selected this sample rate based on the fact that a typical speaking rate for English is roughly 4 syllables per second [16]. 8.5 FPS is roughly twice this frequency and should be sufficient to capture 4 Hz movements.

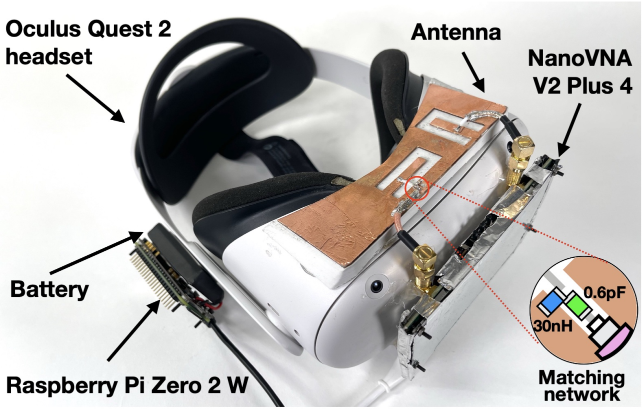

Figure 6: Labeled photo of our Pantœnna prototype showing our final Dual Asymmetric Enhanced Half-Wavelength Antenna design.

Lastly, as our prototype is hand built, there are small variances in the SMA cables, connectors, fixed dielectrics (headset, support wedge), and indeed all of the components. Even in a mass-produced product, there will be small variances that can affect the signal sensitivity. For these reasons, our real-world signals generally do not immediately match our idealized computational simulations (i.e., a perfect VNA, perfect connections, etc.). Instead, we perform a one-time automatic calibration of our VNA affixed to the headset and connected to our antenna, which would happen at the factory in a commercial setting.

4.4 Wireless Operation

As already noted, we tethered our VNA to a laptop over USB to simplify control and power. However, both the Quest 2 headset and VNA can be made battery-powered and wireless, which is what we demonstrate in our Figures and several parts of our Video Figure. More specifically, we run the Quest 2 using its AirLink mode and our VNA is connected and controlled by a Raspberry Pi Zero 2 W powered by a 4.44 Wh battery (Figure 6), which provides several hours of runtime. In this wireless configuration, data is streamed by the Raspberry Pi to a laptop over WiFi for e.g., recording or machine learning.

4.5 Machine Learning Pipeline

Our mouth pose estimation model runs on the aforementioned laptop, though we note in a commercial implementation that our lightweight model could easily run on the processor of the Meta Quest 2.

As explained in Section 2.3, the outputs of our sensing pipeline are four vectors: S11 return loss magnitudes, S11 phase shifts, S21 return loss magnitudes, and S21 phase shifts. Each of these vectors is either 61 or 31 values in length, for a total of 244 values for sensing facial expression or 144 values for sensing speech movements. In addition to using these raw measurements as machine learning features, we perform additional featurization. Specifically, for each vector, we compute the first derivative of the series (61 or 31 features × 4), the difference from the previous frame of data (60 or 30 × 4), standard deviation (1 feature × 4), the coefficients from a 3rd order polynomial fit (4 features × 4), and subtraction of S11 and

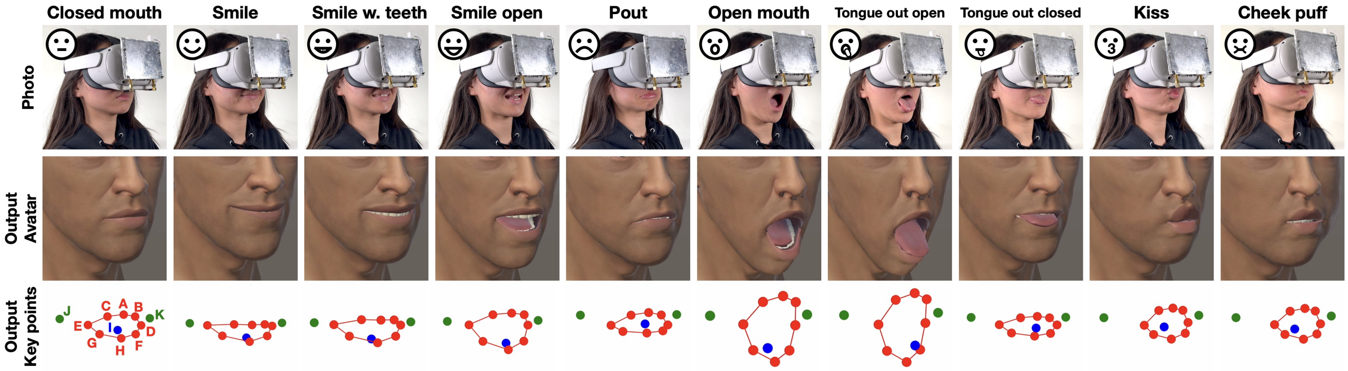

Figure 7: The ten facial expressions requested in our user study (top row) and example output of an avatar (middle row) and mouth keypoints (bottom row; red for lip keypoints, green for cheek, and blue for tongue tip). Left-most mouth keypoints labeled as key for Figure 8.

S21 vectors (61 or 31 features × 2). This yields a total of 870 or 450 features for machine learning.

For continuous mouth pose estimation, we use SciKit Learn’s Extra Tree Regressor [38] with 300 estimators. The outputs of this model are 11 3D mouth keypoints [4, 5], seen in Figures 1 & 7. In our evaluation, we also report expression classification results to facilitate comparison of our system to prior works performing only discrete classification. For this, we use SciKit Learn’s Extra Trees Classifier [37] (number of estimators 300; default parameters otherwise) using the same feature vectors.

4.6 Visualization

In addition to mouth keypoints, we also use rigged 3D heads in Unity (Shieh avatar v2 in Figures 1 & 7) to visualize our tracking results. This head is controlled by 37 SRanipal blendshapes [19], which is also the native output of the VIVE Face Tracker we use for ground truth data capture (discussed later). Please also refer to our Video Figure.

5 EVALUATION

We recruited 12 participants (mean age 26.7, min 20, max 33; 7 identified as female, 5 as male) for a 90-minute user study that paid $30. The width of each participant’s mouth (lip corner to corner) was recorded in the *closed mouth* pose (mean=49.4 mm, SD=3.9), which we used to scale participants’ 3D head model in our software.

We captured two different datasets to evaluate our system’s accuracy. First we capture facial expressions, which tend to be more exaggerated movements held for brief periods of time. Second are mouth movements from speech, which tend to be more continuous and rapid, but smaller in scale. As such, these two datasets are highly complementary. We now describe the details of our study procedure and apparatus.

5.1 Apparatus & Ground Truth Data Capture

Participants wore our proof-of-concept headset, which was tethered to a laptop for experimental control. In VR, participants were shown a study interface that provided visual instructions. During data collection, neither the experimenter nor participants were allowed to view sensor signals or any tracking output, to prevent biasing their behavior. For the duration of the study, participants were seated.

In order to benchmark the accuracy of our system, we needed to ascertain a ground truth mouth pose. The most accurate technologies available today are camera-based systems, which keypoint a high-resolution video stream of a user’s mouth. For this, we use VIVE’s Mouth Tracker camera [68], which can be added to any VR headset as an accessory. The accompanying SDK outputs 37 blendshapes [19] that we use to pose a head and extract keypoints, and which are scaled to match the measurement of our participants. The study apparatus, along with example avatar (middle row) and mouth keypoint output (bottom row) can be seen in Figure 7. We verified the VIVE camera had minimal effect on our system’s sensing. Before data collection began, participants mimicked several mouth poses and said the words “Human-Computer Interaction” to confirm the ground truth camera’s performance, and the experimenter adjusted the camera angle as needed.

5.2 Facial Expression Procedure

We selected 10 mouth-centric facial expressions (Figure 7) drawn from prior work [10–12, 42]. Participants were shown a facial expression in the VR interface and asked to match the mouth pose. Once the requested facial expression was assumed, the experimenter used a key on the laptop to record the timestamp, which we use for later assessment of discrete pose classification accuracy. One round of data collection consisted of all 10 facial expressions being requested in a random sequence. Five rounds formed one session of data capture (i.e., 5 rounds of 10 facial expressions).

Importantly, throughout this procedure, synchronized sensor and ground truth data were captured continuously (5 FPS in this study). Such a procedure helps to capture a wide variety of intermediate poses between the ten terminal facial expressions (10! pairwise combinations). In total, ten sessions of data were captured. After each session was completed, users were asked to remove the headset and take a brief break. Once ready, they re-wore the headset. This

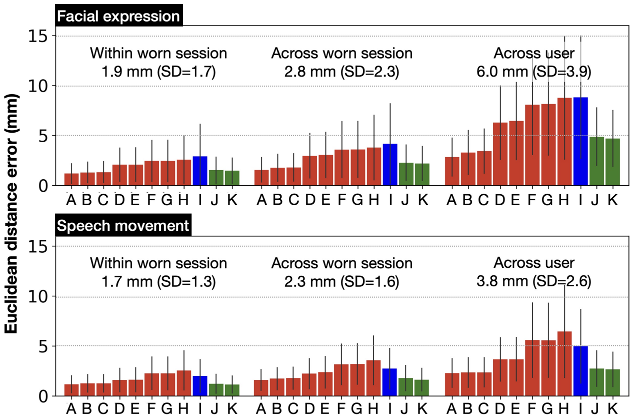

Figure 8: Continuous mouth pose tracking results, reported in mean per-joint position error (MPJPE). Mouth keypoint labels in Figure 7.

adds natural variance to sensor position as is a crucial procedural step for any worn system evaluation.

This procedure yielded 63,480 data points for continuous pose tracking (representing roughly 3.5 hours of data). Of these, 6000 data points had labeled classes (12 participants × 10 worn sessions × 5 rounds × 10 mouth poses) which we use to evaluate discrete mouth pose classification accuracy.

5.3 Speech Movement Procedure

The second stage of data collection focused on mouth movements resulting from speech, which tend to be more rapidly changing than facial expressions. As before, participants saw visual prompts in VR: this time 10 random sentences, shown one at a time, randomly drawn from the CommonVoice dataset [3]. Each worn session contained 5 rounds of 10 random sentences (i.e., 50 sentence utterances per session). Same as earlier, we recorded synchronized sensor and ground truth data. Also matching the previous procedure, participants took off the headset between sessions, adding additional variability and realism to the data.

In total, 5 sessions of data were collected per participant, yielding 174,764 data points for continuous mouth pose tracking (representing roughly 5.7 hours of data). No discrete mouth pose data was collected in this procedure.

6 RESULTS

Using our study data, we explored three training procedures that emulate different user calibration and training schemes. An overview of results can be found in Figures 8 and 9.

6.1 Within-Worn-Session Accuracy

Worn bio-sensing systems (e.g., EMG [35], EIT [76]) generally require re-calibration each time the sensor is worn in order to provide usable accuracies. This obviously offers the highest accuracy, but comes at a significant cost to the user experience, and generally should be avoided. However, we report these results so as to enable comparison to prior work using this "within-worn-session" procedure. As described above, each of our worn sessions (for both facial expression and speech movements) contained 5 rounds of data collection. This means we can use all combinations of 4 rounds for training and 1 round for testing to measure within-worn-session accuracy. Further, we can do this for all worn sessions collected, averaging the results to produce a more reliable estimate of accuracy.

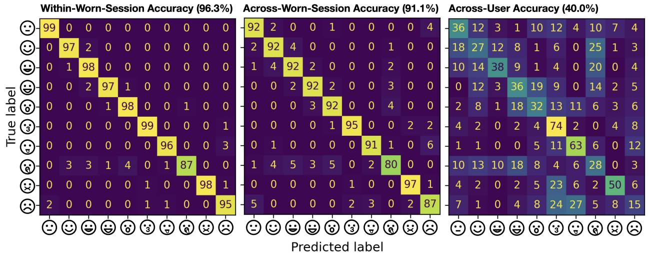

Figure 9: Our system's primary mode is continuous mouth pose estimation. However, to facilitate comparison to prior work, we also computed discrete pose classification accuracy. The classification accuracy is 96.3% for within-worn-session and 91.1% for across-worn-session. We also tested across-user (i.e., "out of the box") accuracy, which was 40.0%.

Overall, we found a within-worn-session tracking accuracy of 1.8 mm (SD=1.5). For our facial expression study data, tracking accuracy was 1.9 mm with a discrete expression classification accuracy of 96.3% (SD=1.8). For our speech movement data, we found an accuracy of 1.7 mm (SD=1.3). See also Figures 8 and 9.

6.2 Across-Worn-Session Accuracy

Given the variability of human mouth morphology, and even how we perform different expressions and move our mouths in speech, it is not unreasonable to imagine that users will have to perform at least some system calibration (especially non-camera methods). To simulate this, we train our model on all-but-one-session of a user's data, using a hold-out round for testing. This procedure is done for all worn session combinations, with results combined. As a reminder, participants removed the headset between sessions, so this simulates calibrated, but across-worn-session use.

Combining both datasets, we found an across-session tracking accuracy of 2.6 mm (SD=2.0). Looking specifically at facial expressions, tracking accuracy was 2.8 mm (SD=2.3) with a discrete expression classification accuracy of 91.1% (SD=3.4). For our speech movement data, we found an accuracy of 2.3 mm (SD=1.6). See also Figures 8 and 9.

6.3 Across-User Accuracy

Ideally, users would not have to provide any calibration data, and could simply wear a headset and mouth tracking would work "out of the box". This is the most challenging train/test arrangement, as human mouth morphology varies tremendously, as does how we express our emotions and speak. For this reason, we believe at least some calibration data would be needed in a consumer system. Nonetheless, we report across-user accuracy by training our models on data from all-but-one participant, and then using a hold-out participant for testing. We can do this for all combinations of users (all combinations, results average).

Across all sessions, we found our system had an overall tracking accuracy of 4.9 mm (SD=3.3). On our facial expression data, tracking accuracy was 6.0 mm (SD=3.9) with discrete expression classification accuracy of 40% (SD=14.9%). For our speech movement data, we found an accuracy of 3.8 mm (SD=2.6). See also Figures 8 and 9.

6.4 User Identification

We also used our study data to test an additional, orthogonal feature of our system: user identification. VR headsets are often shared within a family or group of colleagues or friends, and thus a system that can recognize its current user among a small group has great value. For instance, it could load personalized content (i.e., system settings, game progress, interpupillary distance) or provide automatic content controls for children.

As an initial test of this potential feature, we used our 12 participants as a simulated "family". We trained the model using all-but-one common session of their data (e.g., all participants' sessions 1-9 were used for training, and all session 10 data was used for testing. Then all sessions 1-8 and 10 were used for training, and all participants' session 9 was used for testing, and so on). Class labels were Participant_1 through Participant_12. The model is deemed correct if it correctly predicts the family member (i.e., study participant), and incorrect if it guesses any of the other 11 possible family members. In this simulated family, our system was 99.5% accurate (SD=0.07) at recognizing the participant, significantly exceeding our expectations and meriting future work.

7 DISCUSSION

7.1 Comparison to Prior Work

Table 1 provides an overview of prior work, both discrete mouth pose and continuous tracking systems. A good point of comparison is the recently published C-Face [11] system, which uses ear-worn cameras to capture side-of-face contour changes. Using this arrangement, C-Face offers 2.8 mm of continuous mouth pose tracking accuracy across worn sessions (the same as Pantœnna's 2.8mm error). In comparison to C-Face, Pantœnna seems to excel at smaller mouth movements. For example, *smile* vs. *smile with teeth* does not significantly move the cheeks. C-Face [11] demonstrates 87% classification accuracy for *smile* and 84% classification accuracy for *smile with teeth* (89% classification accuracy over nine facial expressions in total). Pantœnna, on the other hand, is 96.7% and 98.3% accurate on *smile* and *smile with teeth* classes, respectively (and 96.3% accurate on ten facial expressions overall).

We can also compare Pantœnna to other worn, non-camera-based systems. These systems generally report within-worn-session accuracy, so we use this as a comparative metric. We note that such comparisons are only approximate as different participants, pose sets, facial landmarks, and procedures were used. Pantœnna's within-worn-session accuracy was 1.8 mm (SD=1.5), and 1.7 mm specifically for our speech condition. We can compare this result to BioFace-3D, [72] which used EMG and EOG signals to achieve 2.39 mm of mouth tracking error in a similar speech task. For within-session mouth pose classification, the acoustic-driven Interferi [29] system achieved 91.1% accuracy on five mouth poses, while our system demonstrates 96.3% accuracy over ten mouth poses. Interferi [29] also reported across-worn-session accuracy, which dropped to 68.4%. Pantœnna's accuracy drops from 96.3% within-session to 91.1% across-session.

7.2 Privacy

As noted in our Introduction and Related Work sections, camera-based methods, currently the most popular and successful approach, raise significant privacy issues. First off, the high-resolution imagery of the mouth permits the identification of individuals (not only from distinctive face and mouth morphology, but also dental biometrics). Second, the field of view of many vision-based mouth pose systems includes the user's upper torso, capturing an oblique downward view of a user's breasts, which is intrusive. Third, other people and sensitive content can be inadvertently captured in the video stream, such as financial and medical documents.

Ideally, no visual data would be captured, and in this regard, Pantœnna makes a significant contribution. That said, Pantœnna does not fully alleviate privacy concerns. For instance, as we evaluated, Pantœnna can be used to recognize users within small groups (though not with larger populations). It may also be that with enough data, the facial geometry of a user could be reconstructed. Finally, tracking people's mouth and lips could allow for the reconstruction of speech content without audio. Of course, camera-based systems make such attacks considerably easier.

7.3 Cost

Our antenna is made from acrylic and copper tape and has negligible cost. In a commercial system, this would almost certainly be a low-cost PCB antenna, perhaps integrated onto an existing board. Almost all of our prototype's cost comes from a $230 NanoVNA V2 Plus4. This is a sophisticated measurement instrument with capabilities far exceeding our needs. Inexpensive VNAs, such as the NanoVNA-H, can be found at popular online retailers for under $50. This device includes a 2.8" color touchscreen, enclosure, battery, and USB-C connectivity — the actual VNA components likely cost under $20. We also note there continue to be significant advances in single chip VNAs [14, 15, 20], which could allow for exceptionally compact implementations. Finally, it may also be possible to avoid using a full-featured VNA and instead used a fixed set of RF frequencies with dedicated hardware.

8 LIMITATIONS & FUTURE WORK

There are several avenues of future work, which we now briefly discuss. First and foremost among these is the continued reduction in the size of our technique. Our current design uses a small wedge to better orient our antenna's radiation pattern towards the mouth, but which is larger than the antenna itself. In the future, it may be possible to use a different antenna topology or perhaps even an antenna array to beamform, further reducing the volume of our technique and potentially permitting integrations into lightweight, glasses-like form factors. Additionally, although our antenna's directivity helped to reduce interference from the environment and the user's hands, it is still susceptible. In particular, dropping the chin down to the chest detrimentally alters the signal and reduces accuracy (other neck angles appear to be less problematic). We also note that the VNA we used was not capable of capturing S22 data, even though it appears to be a useful and information-bearing signal. It is highly probable that future iterations capable of capturing S22 data would achieve even higher accuracies.

Our prototype system used two separate machine learning models, one for facial expression and another for speech-related mouth movements. This was due to the limited sampling rate of our VNA (discussed in Section 4.3), forcing us to trade off resolution vs. framerate for different use cases. However, importantly, this is not an innate limitation of our technique. Z-Ring [69], for example, utilized a different low-cost VNA and achieved 30 FPS. Transitioning from a general-purpose, wide-bandwidth VNA to specialized hardware could provide an order-of-magnitude improvement in framerate. Even still, it may be useful to have separate mouth pose models for expressions and speech. These could be automatically toggled with, for example, the loudness of audio input.

Finally, we note that the number of participants in our user study is relatively small and tends toward a younger demographic, which is not representative of the whole population. That said, while faces do change with age, the most significant effects we leverage are morphological in nature – e.g., the lips separating to reveal the teeth during a *smile*, the jaw opening to say “ah”, or the tongue sticking out in front of the lips. These are gross facial geometry differences that are true across all ages, and which manifest in our sensor signal.

9 CONCLUSION

We have presented our work on Pantœnna, a new continuous mouth pose sensing method. By not using either cameras or microphones, our approach sidesteps significant issues in privacy, while still offering substantial facial expressivity. We explored both facial expressions and mouth movements during speech. Our user study revealed a mean 3D Euclidean error of 2.6 mm across worn sessions, a train/test condition that most bio-sensing systems find challenging (due to their sensitivity to worn placement). While future work remains, our present Pantœnna proof of concept demonstrates a unique set of pros and cons, which could be combined with other approaches in a multimodal fashion.

ACKNOWLEDGMENTS

We are very grateful to Istvan Szini for his deep expertise and assistance in characterizing the phenomena used in this paper, as well as his help in running the software simulations that informed the design of our system.

REFERENCES

[1] Karan Ahuja, Chris Harrison, Mayank Goel, and Robert Xiao. 2019. Mecap: Whole-body digitization for low-cost vr/ar headsets. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology.* 453–462.

[2] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. CanalSense: Face-Related Movement Recognition System Based on Sensing Air Pressure in Ear Canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17).* Association for Computing Machinery, New York, NY, USA, 679–689. https://doi.org/10.1145/3126594.3126649

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* (2019).

[4] Autodesk. 2022. Softimage User Guide: The motion capture process. (2022). https://download.autodesk.com/global/docs/softimage2014/en_us/userguide/index.html?url=files/face_act_TheMotionCaptureProcess.htm,topicNumber=d30e233562

[5] Autodesk. 2022. Softimage User Guide: What's on Act Panel? (2022). https://download.autodesk.com/global/docs/softimage2014/en_us/userguide/index.html?url=files/face_act_WhatsontheActPanel.htm,topicNumber=d30e232136,hash=WSF468A8297860DD449B69C7615D16EBA9-0006

[6] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. 2020. FitByte: Automatic Diet Monitoring in Unconstrained Situations Using Multimodal Sensing on Eyeglasses. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376869

[7] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: using wearable sensors to detect eating episodes in unconstrained environments. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–20.

[8] Ho-Seung Cha and Chang-Hwan Im. 2022. Performance enhancement of facial electromyogram-based facial-expression recognition for social virtual reality applications using linear discriminant analysis adaptation. *Virtual Reality* 26, 1 (2022), 385–398.

[9] Jin-xiang Chai, Jing Xiao, and Jessica Hodgins. 2003. Vision-based control of 3D facial animation. In *Symposium on Computer animation*, Vol. 2. Citeseer.

[10] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, Francois Guimbretiere, and Cheng Zhang. 2021. Neckface: Continuously tracking full facial expressions on neck-mounted wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–31.

[11] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-face: Continuously reconstructing facial expressions by deep learning contours of the face with ear-mounted miniature cameras. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology.* 112–125.

[12] Victor Chen, Xuhai Xu, Richard Li, Yuanchun Shi, Shwetak Patel, and Yuntao Wang. 2021. Understanding the Design Space of Mouth Microgestures. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21).* Association for Computing Machinery, New York, NY, USA, 1068–1081. https://doi.org/10.1145/3461778.3462004

[13] Yumiao Chen, Zhongliang Yang, and Jiangping Wang. 2015. Eyebrow emotional expression recognition using surface EMG signals. *Neurocomputing* 168 (2015), 871–879.

[14] Hyunchul Chung, Qian Ma, Mustafa Sayginer, and Gabriel M Rebeiz. 2017. A 0.01–26 GHz single-chip SiGe reflectometer for two-port vector network analyzers. In *2017 IEEE MTT-S International Microwave Symposium (IMS).* IEEE, 1259–1261.

[15] Hyunchul Chung, Qian Ma, Mustafa Sayginer, and Gabriel M Rebeiz. 2020. A Packaged 0.01–26-GHz single-chip SiGe reflectometer for two-port vector network analyzers. *IEEE Transactions on Microwave Theory and Techniques* 68, 5 (2020), 1794–1808.

[16] Alan Cruttenden. 2014. *Gimson's pronunciation of English.* Routledge.

[17] CST. 2022. CST Studio Suite 3D EM simulation and analysis software. (2022). https://www.3ds.com/products-services/simulia/products/cst-studio-suite/

[18] Charles Darwin and Phillip Prodger. 1998. *The expression of the emotions in man and animals.* Oxford University Press, USA.

[19] VIVE developers. 2022. Integrating facial tracking with your avatar. (2022). https://developer.vive.com/resources/openxr/openxr-pcvr/tutorials/unity/integrate-facial-tracking-your-avatar/

[20] Analog Devices. 2022. 10 MHz to 20 GHz, Integrated Vector Network Analyzer Front-End. (2022). https://www.analog.com/media/en/technical-documentation/data-sheets/adl5960.pdf

[21] John V Draper, David B Kaber, and John M Usher. 1998. Telepresence. *Human factors* 40, 3 (1998), 354–375.

[22] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384.

[23] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. 2018. Generating talking face landmarks from speech. In *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Guildford, UK, July 2–5, 2018, Proceedings 14.* Springer, 372–381.

[24] Michael J Fagan, Stephen R Ell, James M Gilbert, E Sarrazin, and Peter M Chapman. 2008. Development of a (silent) speech recognition system for patients following laryngectomy. *Medical engineering & physics* 30, 4 (2008), 419–425.

[25] Chris Frith. 2009. Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3453–3458.

[26] Pablo Gallego Cascón, Denys J.C. Matthies, Sachith Muthukumarana, and Suranga Nanayakkara. 2019. ChewIt. An Intraoral Interface for Discreet Interactions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19).* Association for Computing Machinery, New

York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300556

[27] Anna Gruebler and Kenji Suzuki. 2010. Measurement of distal EMG signals using a wearable device for reading facial expressions. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 4594–4597.

[28] Anna Gruebler and Kenji Suzuki. 2014. Design of a Wearable Device for Reading Positive Expressions from Facial EMG Signals. *IEEE Transactions on Affective Computing* 5, 3 (2014), 227–237. https://doi.org/10.1109/TAFFC.2014.2313557

[29] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture sensing using on-body acoustic interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[30] Shoya Ishimaru, Kai Kunze, Koichi Kise, Jens Weppner, Andreas Dengel, Paul Lukowicz, and Andreas Bulling. 2014. In the Blink of an Eye: Combining Head Motion and Eye Blink Frequency for Activity Recognition with Google Glass. In *Proceedings of the 5th Augmented Human International Conference* (Kobe, Japan) *(AH '14)*. Association for Computing Machinery, New York, NY, USA, Article 15, 4 pages. https://doi.org/10.1145/2582051.2582066

[31] S. Kanoh, S. Ichi-nohe, S. Shioya, K. Inoue, and R. Kawashima. 2015. Development of an eyewear to measure eye and body movements. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2267–2270. https://doi.org/10.1109/EMBC.2015.7318844

[32] Blair Kidwell and Jonathan Hasford. 2014. Emotional ability and nonverbal communication. *Psychology & Marketing* 31, 7 (2014), 526–538.

[33] Daehwa Kim and Chris Harrison. 2022. EtherPose: Continuous Hand Pose Tracking with Wrist-Worn Antenna Impedance Characteristic Sensing. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 58, 12 pages. https://doi.org/10.1145/3526113.3545665

[34] Daehwa Kim, Keunwoo Park, and Geehyuk Lee. 2021. Atatouch: Robust finger pinch detection for a vr controller using rf return loss. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–9.

[35] Jonghwa Kim, Stephan Mastnik, and Elisabeth André. 2008. EMG-Based Hand Gesture Recognition for Realtime Biosignal Interfacing. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (Gran Canaria, Spain) *(IUI '08)*. Association for Computing Machinery, New York, NY, USA, 30–39. https://doi.org/10.1145/1378773.1378778

[36] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Zixiong Su, Alex Olwal, Jun Rekimoto, and Thad Starner. 2022. SilentSpeller: Towards Mobile, Hands-Free, Silent Speech Text Entry Using Electropalatography. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 288, 19 pages. https://doi.org/10.1145/3491102.3502015

[37] SciKit Learn. 2022. Extra Tree Classifier. (2022). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

[38] SciKit Learn. 2022. Extra Tree Regressor. (2022). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html

[39] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.

[40] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. Eario: A low-power acoustic sensing earable for continuously tracking detailed facial movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[41] Yang Li and Youngwook Kim. 2016. Classification of human activities using variation in impedance of single on-body antenna. *IEEE Antennas and Wireless Propagation Letters* 16 (2016), 541–544.

[42] Vanessa LoBue and Cat Thrasher. 2015. The Child Affective Facial Expression (CAFE) set: Validity and reliability from untrained adults. *Frontiers in psychology* 5 (2015), 1532.

[43] Jianwen Lou, Yiming Wang, Charles Nduka, Mahyar Hamedi, Ifigeneia Mavridou, Fei-Yue Wang, and Hui Yu. 2020. Realistic Facial Expression Reconstruction for VR HMD Users. *IEEE Transactions on Multimedia* 22, 3 (2020), 730–743. https://doi.org/10.1109/TMM.2019.2933338

[44] Katsutoshi Masai, Yuta Sugiura, Masa Ogata, Kai Kunze, Masahiko Inami, and Maki Sugimoto. 2016. Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 317–326.

[45] Katsutoshi Masai, Yuta Sugiura, Katsuhiro Suzuki, Sho Shimamura, Kai Kunze, Masa Ogata, Masahiko Inami, and Maki Sugimoto. 2015. AffectiveWear: towards recognizing affect in real life. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 357–360.

[46] Denys J.C. Matthies, Chamod Weerasinghe, Bodo Urban, and Suranga Nanayakkara. 2021. CapGlasses: Untethered Capacitive Sensing with Smart Glasses. In *Proceedings of the Augmented Humans International Conference 2021* (Rovaniemi, Finland) *(AHs '21)*. Association for Computing Machinery, New York, NY, USA, 121–130. https://doi.org/10.1145/3458709.3458945

[47] Denys J. C. Matthies, Bernhard A. Strecker, and Bodo Urban. 2017. EarFieldSensing: A Novel In-Ear Electric Field Sensing to Enrich Wearable Gesture Input through Facial Expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1911–1922. https://doi.org/10.1145/3025453.3025692

[48] Ifigeneia Mavridou, James T McGhee, Mahyar Hamedi, Mohsen Fatoorechi, Andrew Cleal, Emili Ballaguer-Balester, Ellen Seiss, Graeme Cox, and Charles Nduka. 2017. FACETEQ interface demo for emotion expression in VR. In *2017 IEEE virtual reality (VR)*. IEEE, 441–442.

[49] Lucie Ménard. 2015. Multimodal speech production. *The Handbook of Speech Production* (2015), 200–221.

[50] Meta. 2020. Meta Quest 2: Immersive All-In-One VR Headset. (2020). https://www.meta.com/quest/products/quest-2/

[51] Meta. 2021. Oculus Lipsync Unity. (2021). https://developer.oculus.com/downloads/package/oculus-lipsync-unity/

[52] Meta. 2022. Face Tracking for Movement SDK for Unity. (2022). https://developer.oculus.com/documentation/unity/move-face-tracking/

[53] Meta. 2022. Horizon Worlds: Virtual Reality Worlds and Communities. (2022). https://www.meta.com/horizon-worlds/

[54] Patrick Mussel, Anja S Göritz, and Johannes Hewig. 2013. The value of a smile: Facial expression affects ultimatum-game responses. *Judgment and Decision Making* 8, 3 (2013), 381–385.

[55] NanoVNA. 2023. NanoVNA V2 Official Site. (2023). https://nanorfe.com/nanovna-v2.html

[56] Kyle Olszewski, Joseph J Lim, Shunsuke Saito, and Hao Li. 2016. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–14.

[57] Gilles Pourtois and Monica Dhar. 2012. Integration of face and voice during emotion perception: is there anything gained for the perceptual system beyond stimulus modality redundancy? In *Integrating face and voice in person perception*. Springer, 181–206.

[58] Ville Rantanen, Hanna Venesvirta, Oleg Spakov, Jarmo Verho, Akos Vetek, Veikko Surakka, and Jukka Lekkala. 2013. Capacitive Measurement of Facial Activity Intensity. *IEEE Sensors Journal* 13, 11 (2013), 4329–4338. https://doi.org/10.1109/JSEN.2013.2269864

[59] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando De la Torre, and Yaser Sheikh. 2021. Audio-and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 41–50.

[60] Lawrence D Rosenblum. 2008. Speech perception as a multimodal phenomenon. *Current directions in psychological science* 17, 6 (2008), 405–409.

[61] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. The tongue and ear interface: a wearable system for silent speech recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers*. 47–54.

[62] Javier San Agustin, John Paulin Hansen, Dan Witzner Hansen, and Henrik Skovsgaard. 2009. Low-Cost Gaze Pointing and EMG Clicking. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 3247–3252. https://doi.org/10.1145/1520340.1520466

[63] T. Scott Saponas, Daniel Kelly, Babak A. Parviz, and Desney S. Tan. 2009. Optically Sensing Tongue Gestures for Computer Input. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology* (Victoria, BC, Canada) *(UIST '09)*. Association for Computing Machinery, New York, NY, USA, 177–180. https://doi.org/10.1145/1622176.1622209

[64] Jaemin Shin, Seungjoo Lee, Taesik Gong, Hyungjun Yoon, Hyunchul Roh, Andrea Bianchi, and Sung-Ju Lee. 2022. MyDJ: Sensing Food Intakes with an Attachable on Your Eyeglass Frame. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[65] SPEAG. 2022. SPEAG, Schmid & Partner Engineering AG. (2022). https://speag.swiss/

[66] Katsuhiro Suzuki, Fumihiko Nakamura, Jiu Otsuka, Katsutoshi Masai, Yuta Itoh, Yuta Sugiura, and Maki Sugimoto. 2016. Facial Expression Mapping inside Head Mounted Display by Embedded Optical Sensors. In *Adjunct Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 91–92. https://doi.org/10.1145/2984751.2985714

[67] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions On Graphics (TOG)* 36, 4 (2017), 1–11.

[68] VIVE. 2021. VIVE Facial Tracker. (2021). https://www.vive.com/us/accessory/facial-tracker/

[69] Anandghan Waghmare, Youssef Ben Taleb, Ishan Chatterjee, Arjun Narendra, and Shwetak Patel. 2023. Z-Ring: Single-Point Bio-Impedance Sensing for Gesture, Touch, Object and User Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[70] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR Facial Animation via Multiview Image Translation. *ACM Trans. Graph.* 38, 4, Article 67 (jul 2019), 16 pages. https://doi.org/10.1145/3306346.3323030

[71] Lihang Wen, Jianlong Zhou, Weidong Huang, and Fang Chen. 2021. A survey of facial capture for virtual reality. *IEEE Access* 10 (2021), 6042–6052.

[72] Yi Wu, Vimal Kakaraparthi, Zhuohang Li, Tien Pham, Jian Liu, and Phuc Nguyen. 2021. BioFace-3D: Continuous 3d Facial Reconstruction through Lightweight Single-Ear Biosensors. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking* (New Orleans, Louisiana) *(MobiCom '21)*. Association for Computing Machinery, New York, NY, USA, 350–363. https://doi.org/10.1145/3447993.3483252

[73] Bin Xu, Yang Li, and Youngwook Kim. 2017. Classification of finger movements based on reflection coefficient variations of a body-worn electrically small antenna. *IEEE Antennas and Wireless Propagation Letters* 16 (2017), 1812–1815.

[74] Zihan Yan, Yufei Wu, Yang Zhang, and Xiang'Anthony' Chen. 2022. Emoglass: an end-to-end ai-enabled wearable platform for enhancing self-awareness of emotional health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[75] Qiao Zhang, Shyamnath Gollakota, Ben Taskar, and Raj P.N. Rao. 2014. Non-Intrusive Tongue Machine Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2555–2558. https://doi.org/10.1145/2556288.2556981

[76] Yang Zhang and Chris Harrison. 2015. Tomo: Wearable, Low-Cost Electrical Impedance Tomography for Hand Gesture Recognition. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 167–173. https://doi.org/10.1145/2807442.2807480

[77] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. 2018. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–10.