IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

SIREN: Underwater Robot-To-Human Communication Using Audio*

Michael Fulton[1], Junaed Sattar[1], and Rafa Absar[3]

***Abstract*—In this paper we present SIREN: a novel audio-based communication system for underwater human-robot interaction. SIREN utilizes a surface transducer to produce sound by vibrating the frame of an underwater robot, essentially turning the robot's outer surface into the vibrating membrane of a speaker. We employ this hardware in two forms of robot-to-human communication: synthesized text-to-speech (TTS-sonemes) and synthesized musical indicators (Tone-sonemes). To profile the system's capabilities with respect to underwater communication, we perform a substantial in-person human study with 12 participants. In this study, participants were trained on the use of one of the previously mentioned audio communication systems. Participants were then asked to identify the communication from their system in a pool at various distances. This study's results demonstrate that sound is a viable method of underwater communication. TTS-Sonemes outperform Tonal-Sonemes at close distances but fail at further distances, while Tonal-Sonemes remain recognizable as the distance to the robot increases.**

***Index Terms*—Human-Robot Collaboration; Marine Robotics; Field Robots**

I. Introduction

EARTH'S water resources are a critical part of the lives of every organism on the planet. Our oceans, lakes, and rivers are home to shipping routes, oil wells, pipelines, internet cabling, archaeological sites of interest, and ecosystems that are critical to the ecological stability of the planet. Divers around the world conduct a vast and varied set of tasks in these environments, studying naturally occurring phenomena and constructing or maintaining synthetic systems. For over seventy years, scientists have been developing autonomous underwater vehicles (AUVs) for the purpose of aiding humans in all of the critical work that is done underwater. In recent years, AUVs have spread into numerous application areas including the development of collaborative AUVs, which work alongside divers rather than replace them. This growing field has necessitated the development of new methods for underwater human-robot interaction (UHRI) [1]–[4].

High-fidelity interaction is necessary for collaborative work. However, underwater environments are adversarial to standard forms of communication, making UHRI particularly challenging. Gestural communication has naturally become one of the most common methodologies for human-to-robot (H2R) communication in underwater environments [4], taking advantage of the already ubiquitous gestural languages used by divers. What then of the inverse task, robot-to-human (R2H) communication? Robot-to-human communication has largely been dominated by the use of digital displays [4], which are difficult to read at many distances or angles. To a lesser extent, this type of communication has been achieved by the use of emitted light and robotic body language gestures. Aside from power/status-indicating tones, audio communication has not been significantly explored for underwater robots. Sound travels well through water, but producing and comprehending it is challenging [5]. Most commercially available speakers are not designed for vibrating water rather than air, while underwater-compatible speakers tend to be quite expensive, and incompatible with small AUVs. Additionally, human auditory processing is not well suited for comprehending sound underwater, leading to confusion and garbling of complex signals such as speech. Due to these confounding challenges, audible communication from robots to humans underwater is largely unexplored.

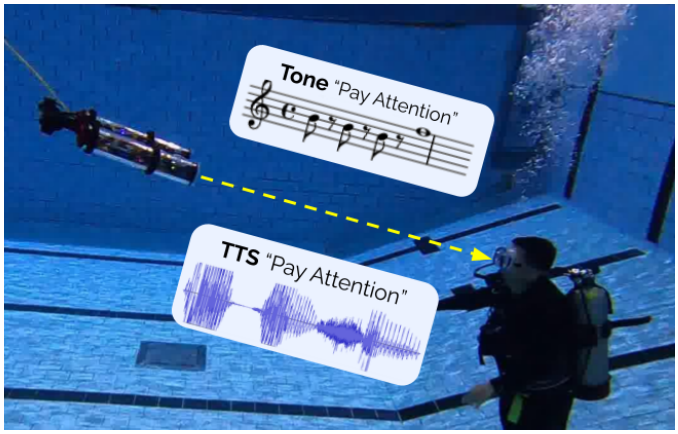


Fig. 1: A depiction of the two types of sonemes that SIREN can produce, asking a diver for their attention.

We present a novel audio-based system for underwater robots, named SIREN (**S**ound **I**ndicators via **R**esonance **E**xciters u**N**derwater). In the following sections, we first discuss the background of this work, exploring the types of robot-to-human communication which have been developed thus far. Next, we present the hardware and software used to create the audible communication indicators of SIREN, which we refer to as **sonemes**. We expand upon the design of these sonemes in the next section, defining two types of

Manuscript received: February 2nd, 2023; Revised: June 1st, 2023; Accepted: July 7th, 2023.

This paper was recommended for publication by Editor Angelika Peer upon evaluation of the Associate Editor and Reviewers' comments.

*Research supported by the NSF grants #00074041 and IIS-2220956.

[1]Authors associated with the University of Minnesota Twin Cities, Department of Computer Science and Engineering.

[2]Author associated with the Metro State University, St. Paul, Minnesota, Department of Computer Science and Cybersecurity.

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

sonemes: synthetic speech (TTS-sonemes) and musical tone indicators (Tone-sonemes). To evaluate these sonemes for use in communication, we perform a substantial in-person human study with a total of 12 participants, in which swimmers identified sonemes at a variety of distances. While a study population of twelve is not considered large in many fields, this is the second largest UHRI study conducted in an underwater environment, due to the logistical challenges of performing such studies. The results from this study reveal the effectiveness of audible communication underwater at close range, with some viability at distance.

Defintion: Soneme – A soneme is a sound intentionally produced by a robot for interaction with a human. These sounds have also previously been referred to as *earcons* [6]or *semantic-free utterances* [7]. To continue parity with our previous work on motion and light-based communications (termed *kinemes* [2] and *lucemes* [8], respectively), we refer to robotic audio communication phrases as *sonemes*. The word soneme is derived from the Latin *sonus* (meaning sound) and the suffix *eme* used for phonemes and cheremes, fundamental parts of audible and gestural languages.

II. Background

A. Underwater Human-Robot Interaction

For the task of robot-to-human communication underwater, the dominant method has long been displaying text on a screen [1], [9]. The size of these screens varies, but their performance is relatively similar: complex and high-density information can be easily communicated, but viewing angles and distances are poor. Wrist-mounted displays and tablets can address the viewing angle/distance issue, but these devices have limited commercial availability. Additionally, requiring a diver to carry a tablet or smartwatch adds to their equipment burden and limits the number of people who can communicate with the AUV to those with a device [10]. To expand the in-built communication capabilities of AUVs, we have previously proposed the use of motion [2], [3], [11] which is much more resistant to distance and orientation changes. Beyond motion, we also have expanded upon early, simple use of emittedight for communication [8], [12] , both of which have improved performance at increased distance and non-standard orientations. However, a problem still remains: what if no one is looking at the robot or the visibility is negligible? In those cases, any kind of communication based on the visibility of the robot will fail. For this reason, we turn our attention to sound, which passes well through water, can be omnidirectional, and requires no visibility.

B. Sound-Based Robot Communication

Sound is a common vector for robot-to-human communication, exploiting what Andrea Bonarini [13] calls the "Hearing Channel". The study of sound-based communication is broad, with a wide variety of applications of sound being investigated. A great deal of work has focused on synthetic speech for robots [14]–[16], particularly the ability of a robotic voice to convey emotion [17], [18]. Another aspect of robotic speech

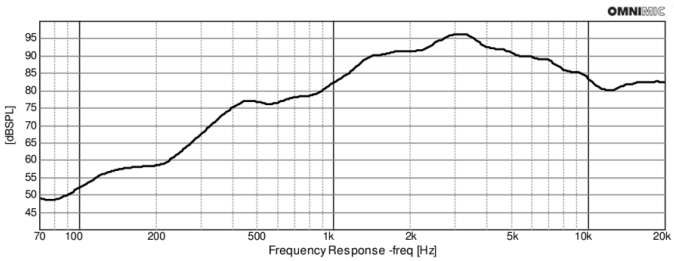


Fig. 2: Frequency response of the Dayton Audio DAEX25W-8, from the provided specification sheet.

that has been well-studied is the effect of different types of voices on robotic acceptance and social behavior [19]. However, not all robotic sound is speech. A number of works have explored the effect that consequential sound – sound that the robot makes as a consequence of its operation – has on the perception of robotic operations and interactions [20], [21]. Further investigation has also considered the addition of artificial sound to consequential sound and the way that it affects perceptions of safety and capability [22]. Sound has also been used as a nonverbal signal to improve human perception of a robot's location [23]. While the use of nonverbal sound is less common in the field than speech, it has been applied to topics such as emotion and intention expression [24]. Similar works have ventured into generative methodologies, where properties of a robot's internal state or emotional cues are directly input into a sound synthesis engine to control different parameters [25]. The design of these types of nonverbal sound communication inspired the tonal sonemes we describe later in this work, though our indicators are pre-defined.

C. Sound-Based Communication Underwater

To our knowledge, this is the first system for sound-based AUV-to-human communication beyond startup tones. However, audio is frequently used for diver-to-diver and ship-to-diver communications. Diver recall systems are common in commercial diving, ranging from an underwater speaker to a simple wrench tapping a tank. More complexly, a variety of wired and wireless systems allow for diver-to-ship communication and between-diver communication. These systems are relatively affordable and commercially available, but we are interested in providing device-less audio communication, which does not require the divers to carry equipment. However, if one was adapting a robot to be used by a group of divers who use such audio communication devices, integrating a modem into the robot and outputting a high-quality text-to-speech system would likely be effective.

III. SIREN System Design

A. Hardware Design

Two pieces of hardware are required for SIREN: a transducer/exciter to vibrate against the frame of an AUV and an amplifier to drive the said transducer. The amplifier also requires a source of audio input, but as AUVs have onboard computers, this is not considered part of the required hardware. SIREN utilizes the DAEX25W-8 [26] waterproofed surface

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

transducer produced by Dayton Audio, which is IP67 rated, having undergone one hour of immersion in one meter of water. Our transducer has been immersed to two meters for extended periods of time, but if a deeper depth rating is required, many other exciters can be acquired and affixed inside an AUV's shell, circumventing the need for greater waterproofing. The exciter has a relatively strong frequency response at all frequencies, though it is most capable between 1k-10k Hz, as seen in Figure 2. However, the frequency response of our system is likely to be different, given that the material the exciter is attached to will have the greatest bearing on its ability to produce certain frequencies. We power our exciter with a small dual 10W amplifier from Parts Express, but any amplifier which fits in the available space and adheres to the power requirements of the exciter used would be appropriate. The cost for both parts totals 35 USD, making SIREN hardware an affordable addition to even the most economical of AUVs.

B. Software Design: Reconfigurable Sound

The sonemes of the SIREN device are produced dynamically by a software module for PROTEUS, our UHRI software system for the Robot Operating System (ROS). PROTEUS loads XML definitions of sonemes and then builds ROS services to trigger them as needed. We wished to explore both synthetic speech and more abstract audio cues as a method of communication, so our software has two modes: **Tonal-Sonemes** and **TTS-Sonemes**.

1) Tonal-Sonemes: For tonal sonemes, PROTEUS expects an XML definition file that contains two things: a system configuration section defining the various waveforms to be used, and a set of soneme definitions. After parsing the definition file, the PROTEUS Tonal-Sonemes server uses a package called *tones* [27] to synthesize polyphonic music.

2) TTS-Sonemes: In the case of TTS-Sonemes (Text-To-Speech Sonemes), PROTEUS expects an XML definition file with a system configuration section specifying the voice, language, and volume to be used, as well as a set of soneme definitions. Once these definitions are parsed, a python package named *voxpopuli* [28] is used to interface with Espeak [29] and MBROLA [30]. Espeak is responsible for parsing the text into a list of phonemes, which MBROLA then synthesizes into audible speech. These modules were selected over more high-quality text-to-speech software could since they do not require an internet connection or GPU.

IV. Sonemes: Meaningful Sound Symbols

With the hardware and software for SIREN defined, we turn to specifying the sonemes of the system. Before defining sonemes however, we must answer a fundamental question: what might the robot need to say in an interaction?

A. Defining AUV Language Symbols

To create a language of robot communication phrases, we turned our attention to the sign languages in broad use among divers. There are many versions of diver sign language, and most divers have picked up further signs for specific situations they encounter in their dives. By considering instructional material for divers in training (*e.g.*, PADI manuals), we can find a common set of useful signs. The meanings of these signs are shown in Figure 3, along with our next step. After combining signs with the same meaning, we clustered these signs at two different levels. The first level, represented by the circles around signs in Figure 3, was common concepts, such as the cluster of signs which all refer to some kind of environmental danger. The second level, represented by the colors of circles and items, was the relevance of a concept to AUV communication. For instance, concepts relating to diver issues with air or bodily states such as trouble clearing ears are not relevant for an AUV to express. From this point, selection of symbols was simple: any concept cluster relevant for an AUV to communicate was included in the language definition. Some symbols were added, such as *Wait* CMD_S and others were adapted from clusters considered not relevant for AUVs, such as $Malfunction_S$ from the Non-Air Self Problem cluster.

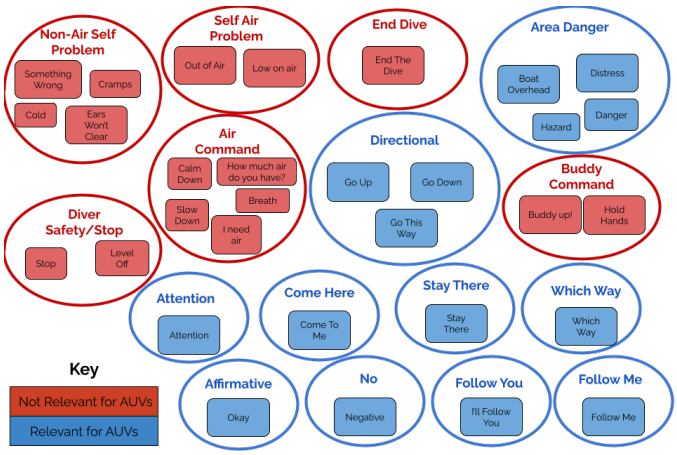


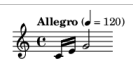
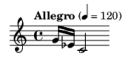
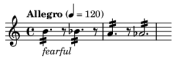
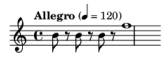
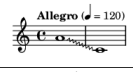
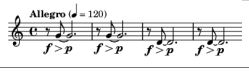
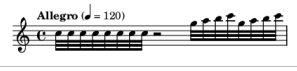
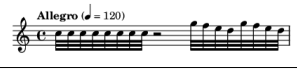
Fig. 3: Clustering of scuba diver sign language symbols, sourced from instructional scuba materials.

B. Soneme Design

With the list of possible sonemes created in this manner, the next step was defining the audio for each soneme, which can be seen in Figure 4 and in the accompanying video.

1) Tonal Sonemes: For tonal sonemes, a variety of techniques were used. Firstly, any soneme with a negative implication (*e.g.*, $Danger_S$, $Negative_S$) was defined to use notes from a minor chord. Additionally, sonemes with positive meanings were designed to sound more cheery or energetic. An effort was made to begin most sonemes or groups of sonemes with unique notes, to reduce overlap between the initial notes of sonemes as much as possible. In addition, the sonemes which did have common start notes were designed to be as distinct as possible from one another other than their starting note, to further avoid confusion. Another design choice was the selection of waveforms used for various sonemes. Sonemes related to commands or information were rendered using a square wave-based tone generator, while directionally related sonemes used a triangle wave.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Soneme\Meaning	Tone Version	TTS Version
Affirmative$_S$ Yes, Okay.		"Yes."
Negative$_S$ No.		"No."
Danger$_S$ Danger in the area.		"Danger nearby!"
Attention$_S$ Pay attention to AUV.		"Pay attention!"
Malfunction$_S$ Internal malfunction.		"I'm experiencing an issue."
Wait CMD$_S$ Waiting for instructions.		"I'm waiting for a command!"
Go Left$_S$ Go left/AUV going left.		"Go left."
Go Right$_S$ Go right/AUV going right.		"Go right."

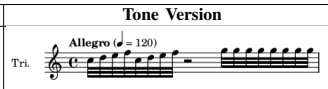
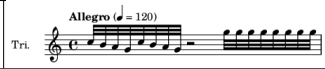
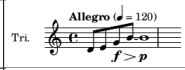
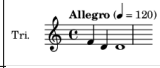
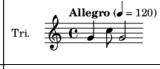
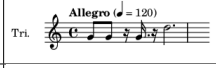
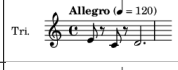
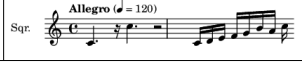
Soneme\Meaning	Tone Version	TTS Version
Go Up$_S$ Go up/AUV going up.		"Go up."
Go Down$_S$ Go down/AUV going down.		"Go down."
Which Way$_S$ Asking for directions.		"Which way are we going?"
Stay$_S$ Stay where you are.		"Stay here."
Come Here$_S$ Come to the AUV.		"Come to me."
Follow Me$_S$ Diver can follow AUV.		"Follow me."
Follow You$_S$ AUV will follow diver.		"I'm going to follow you."
Battery Level$_S$ Battery level is...		"*N*% battery remaining."

Fig. 4: Selected sonemes, with both Tone and TTS versions. See the accompanying video for recordings of each soneme.

2) Text-To-Speech Sonemes: The design of text-to-speech sonemes was simpler, consisting only of selecting an English phrase to communicate the meaning of the soneme. However, some phrases were intentionally lengthened to increase the amount of time that the sound would be audible, and others were modified to avoid confusion with other sonemes.

C. Version Selection Survey

With these design goals in mind, four versions of each soneme were created: two options for TTS-Sonemes and two options for Tonal-Sonemes. These versions were then demonstrated to a small internal focus group comprised of other AUV researchers and non-experts.. While the various TTS phrases were all offered with the same voice, participants were also asked to select one of four voices available from MBROLA by listening to a sample sentence produced by each voice. Based on input from this survey, a final set of sonemes for each version of SIREN was selected by choosing the most popular option for each version, with some designer's discretion in the case of ties. The voice used for TTS-Sonemes was also selected using majority opinion.

V. Human Study: Perception of Sonemes

As the purpose of SIREN is to communicate with divers underwater, the only route to evaluating the effectiveness of the system is to conduct a human study with participants listening to SIREN underwater. Studies of this nature are challenging to administer, as finding and training participants can be time-consuming, costly, and difficult due to low pool availability. The following sections describe the human study which we conducted to evaluate the effectiveness of our SIREN device and the two versions of the sonemes developed for it. This study was approved as human research by the University of Minnesota's Institutional Review Board (reference number: 00016705).

A. Study Design

The study of SIREN efficacy was based on the success of trained participants at recognizing various sonemes underwater. After recruiting participants and training them to a predefined level of competence in recognizing sonemes, we asked them to identify those same sonemes in a pool environment. Participants were randomly assigned to the TTS-Sonemes or Tonal-Sonemes conditions and were trained and tested only on that version using a between-subjects experimental design. Thirty-eight people initiated an intake survey, with fourteen completing study procedures. Unfortunately, the data for two participants was contaminated due to technical issues, yielding twelve total participants' data being included in the analysis.

B. Study Procedures

The study is comprised of three steps: Participant recruitment and training, pool study sessions, and the debrief stage.

1) Participant Recruitment & Training: Participants were recruited from the University of Minnesota by emails, publicly posted fliers, in-classroom announcements, and word of mouth. Participants who were over the age of 18, were not deaf or hard of hearing, able to swim independently, and had not taken part in previous UHRI studies administered by the authors were allowed to complete the intake survey, which collected consent and demographic information in its first part. In the second part of the survey, participants were shown videos with recordings of sonemes in their randomly selected SIREN version. They were taught $3-5$ sonemes at a time, quizzed on their meanings, and then finally given a competency test after learning all sonemes. Only participants who completed the survey and correctly identified at least 12 of the 16 sonemes were able to continue in study procedures. Fourteen participants achieved both requirements (most of the remaining surveys were simply abandoned) and were asked to complete an audiometry test profiling their hearing ability, then schedule a time to complete their pool session.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

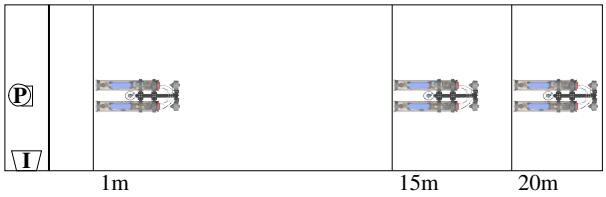


Fig. 5: Experimental setup, featuring LoCO at the three possible distances, a participant (**P**), and the water inlet (**I**).

2) Pool Study Sessions: Three total sessions were administered to collect data from all 14 participants. After participants arrived, study staff explained the session process and read off the list of soneme meanings before beginning. The sounds for each soneme were not played to participants at this point. Once they had been prepared, participants entered the pool and were asked to identify sonemes played by the SIREN device attached to the LoCO AUV [31]. Participants, swimming unequipped, submerged themselves, listened, then surfaced and reported the meaning of the soneme they had heard, along with their confidence in their answer on an ordinal scale from 0 to 10. Sonemes were demonstrated in a randomized order, first at a distance of one meter, then at a further distance of either fifteen or twenty meters. Other research activities were conducted at the same time, resulting in some ambient noise in the pool. Additionally, a water inlet was located to the right of the participants, adding further background noise.

3) Debrief Stage: After completing their pool session, each participant was asked to complete a small debrief survey. This survey uses a modified Godspeed questionnaire [32] to measure attitudes about the AUV, and the NASA Task Load Index [33] survey to measure participant effort and stress during the soneme identification task. Once a participant completed all of their study procedures, they were provided with a $15 Amazon gift card. Once all participants had been enrolled, a $50 gift card was given to a random participant.

C. Analysis of Data

1) Rating and Reliability: While performing pool sessions, participant answers on the meaning of sonemes and their confidence in said answers were recorded. Additionally, the time from the beginning of a soneme to the beginning of a participant's answer was documented. Three independent raters were asked to read through the recorded participant answers and rate the correctness of each answer from one to one hundred. None of the authors of this article served as raters. To determine the level of agreement between raters, Fleiss's κ [34] was calculated to be $\kappa = 0.795$, which is typically understood to indicate a good level of agreement between raters. After determining this, raters' scores were averaged to create the final correctness score.

2) Metrics: When considering the effectiveness of SIREN, we utilize four metrics that have been beneficial in similar analyses in the past [2], [8]:

- **Accuracy:** The average of the rater's correctness scores, which indicates how accurately a participant has identified a soneme.

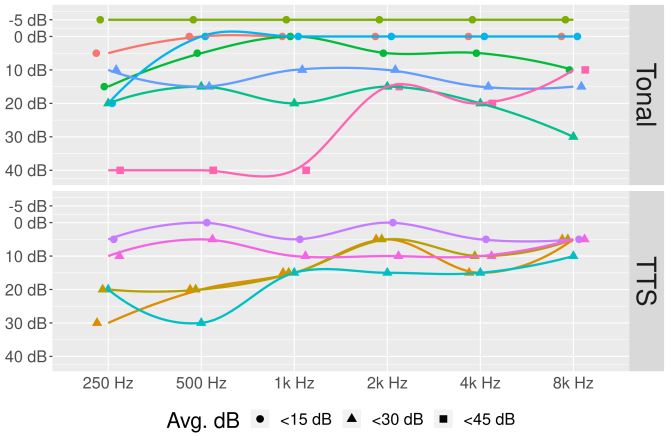


Fig. 6: Audiometry data, expressed as the decibel level required by each participant to hearing the given frequency.

- **Confidence:** A value from 0 to 10, reported by participants, indicated their confidence in their answers.
- **Operational Accuracy:** The same values as accuracy, but only considering answers with a confidence ≥ 6.
- **Time To Answer:** The time from the beginning of a soneme to the beginning of a participant's answer.

3) Statistical Methods: The two metrics analyzed for statistically significant effects are accuracy and operational accuracy. Prior to analyzing our data further, assumptions of statistical tests must be considered. Most parametric statistical tests assume a normal distribution of data. A Shapiro-Wilk [35] test was performed on soneme recognition accuracy $W = 0.71$, $p < .001$, indicating that that data is **not** normally distributed, thus in all analyses, we use non-parametric tests, such as Spearman's correlation [36], Kruskal-Wallis H-tests [37], and Wilcoxon Rank Sum tests [38]. All tests are performed with a significance of $\alpha = 0.01$.

4) Removal of Two Participants: While fourteen people completed all study procedures, two participants had significant issues in their pool sessions which led to corruption or loss of data. Both participants were compensated equally to other participants, but their data is not included .

5) Explanatory Power of Results: The sample size for this study is quite small, with only 12 participants' responses evaluated, leading to a total of 432 observations of sonemes, across two groups (in terms of soneme type) and three groups (in terms of distance). Thus, the statistical testing presented in the following section has limited explanatory power. It cannot be assumed that the results of our testing will hold for any arbitrary population sample, given both our sample size and the number of confounding variables (background noise, participant hearing, etc.). Thus, we urge readers to consider both the significance level and effect size for each test reported and to focus on the bigger picture: the successful demonstration that both soneme types can effectively communicate underwater and the differentiation between the two at different distances.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

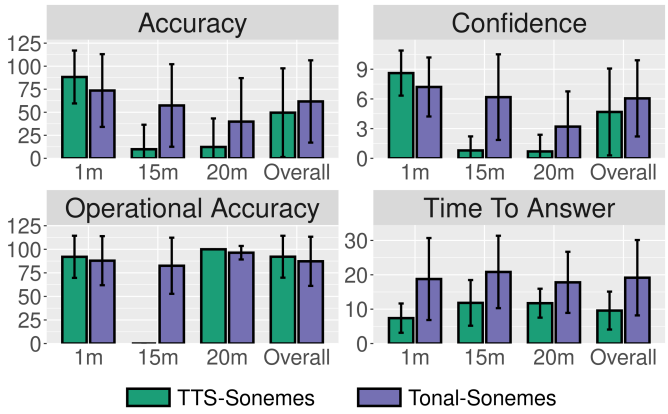


Fig. 7: Metrics for TTS-Sonemes and Tonal-Sonemes.

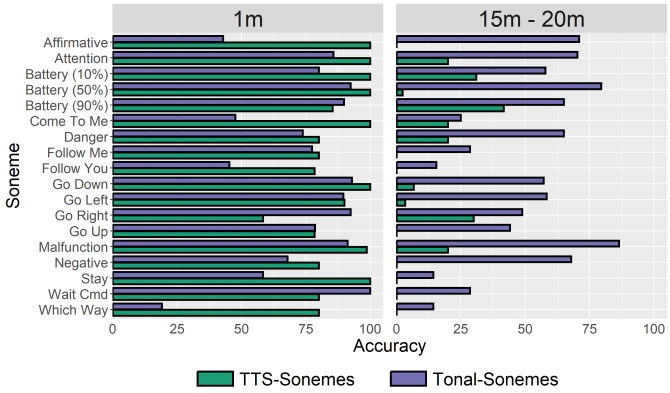


Fig. 8: A comparison of per-soneme accuracy for SIREN at 1m and averaged between 15 and 20m.

VI. Results

A. Population

Our population is fairly small, with seven participants testing Tonal-Sonemes and five testing TTS-Sonemes, though it should be noted that this is an average size for a UHRI human study, which are more logistically challenging than other forms of human evaluations. Approximately half the participants of each condition were tested at distances of 1m and 15m, with the rest tested at 1m and 20m. Eleven participants were between the ages of 18 and 24, with one participant between 35 and 44. Ten participants self-identified as male, with one identifying as female and the last identifying as non-binary/third gender. When asked if they had experience with robots, seven participants answered in the affirmative, and when asked if they had ever been scuba diving, four said they had. No participants self-identified as deaf or hard of hearing and audiometry data showed varying levels of hearing capability, with some falling into the level of mild hearing loss, as can be seen in Figure 6.

B. Internal Validity Tests

The primary situations which threaten the internal validity of this study are training-related. Does the duration of participant training, the level of competency after training, or the time between training and testing have a significant effect on accuracy? Because of the non-normal distribution of accuracy data, non-parametric tests are required, in this case, Spearman's rank correlation test [36] was used. No significant correlation is present between a participant's score on the training test and their accuracy, $r(10) = 0.25, p = 0.44$. Testing the correlation between the time taken during education and accuracy, no significant correlation was found, $r(10) = -0.15, p = 0.65$. Lastly, no significant correlation was found between the accuracy a participant achieved and how long prior to their pool session they had completed their training, $r(10) = -0.07, p = 0.82$. Finally, we consider the effect of a participant's hearing on their accuracy. We first average the decibel level required to hear all tested frequencies from the audiometry test, then assess correlation with participant accuracy (at distances greater than 1m) using Spearman's rank correlation. No significant correlation was found between a participant's hearing ability and accuracy, $r(10) = -0.09, p = 0.77$.

C. Overall SIREN Efficacy

In our pool study, SIREN was demonstrated to be an effective system for AUV-to-diver communication. The metrics previously discussed are presented in Figure 7, with separation between the TTS and Tonal versions, and every metric reported for the three test distances as well as overall. Both versions of sonemes achieved accuracy $\geq 50\%$ overall, with accuracy at 1m test distances $\geq 70\%$. While there is no accepted standard for considering an AUV-to-human communication system "field-viable", these numbers, combined with the operational accuracy of both types of sonemes being $\geq 80\%$ overall indicate that sonemes can be understood underwater. Additionally, the average time to answer values of 5-12 seconds indicate that the time required to understand the system is not significantly more than other systems we have previously evaluated [2], [8].

D. Should We Use TTS-Sonemes or Tonal-Sonemes?

A Kruskal-Wallis test [37] showed that the type of soneme (Tonal or TTS) had no effect on overall accuracy, $H(1) = 5.30, p = 0.02$. However, when considering soneme identification at specific distances, we find a statistically significant difference in accuracy favoring TTS-Sonemes at 1m, $H(1) = 11.57, p < 0.01$, with a small effect size ($\eta^2 = 0.0494$), calculated using the H-statistic [39].. A Kruskal-Wallis test performed on accuracy at 15m and 20m combined also shows a significant effect, $H(1) = 41.55, p < 0.01$, with a large effect size ($\eta^2 = 0.190$), where Tonal-Sonemes lead in accuracy. The effect at 1m was small ($\eta^2 = 0.0494$), while the effect at further distances is large ($\eta^2 = 0.190$). The results of these statistical tests provide a strengthened version of the observation that can be made from Figure 7: TTS-Sonemes are slightly better at 1m while Tonal-Sonemes far outperform TTS-Sonemes at 15m and 20m.

Participant Effort and Stress: In the debrief stage, participants were asked to complete a NASA Task Load Index survey [33]. Kruskal Wallis tests found no statistically significant differences between the answers of participants in the Tonal-Sonemes condition and the TTS-Sonemes condition. In

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

	Tonal-1m	Tonal-15m	Tonal-20m	TTS-1m	TTS-15m
Tonal-15m	0.033	-	-	-	-
Tonal-20m	> 0.001	0.0535	-	-	-
TTS-1m	0.0034	> 0.001	> 0.001	-	-
TTS-15m	> 0.001	> 0.001	0.0019	> 0.001	-
TTS-20m	> 0.001	> 0.001	0.0179	> 0.001	0.7838

TABLE I: The effect of combined condition/distance on soneme accuracy, via pairwise Wilcoxon rank sum tests.

particular, no effect was found on the participant's reported effort, $H(1) = 0.077, p = 0.781$, or frustration, $H(1) = 0.697, p = 0.404$, two areas of major concern.

E. How Does Distance Affect SIREN?

Our findings in the previous section indicate that distance impacts the accuracy of sonemes. This is confirmed with a Kruskal-Wallis test, showing a significant difference in accuracy depending on distnace, $H(2) = 107.75, p < 0.01$, a large effect ($\eta^2 = 0.247$). Furthermore, performing pairwise comparisons using Wilcoxon rank sum testing [38] with Holm-Bonferroni p-value adjustment [40] reveals that there exist significant effects between the $1m$ distance and the others distances, but no significant effects are present between 15m and 20m. By creating a categorical variable combining soneme type and distance (*i.e.*, tts_1m, tone_15m), we can consider the interactions between these variables. A Kruskal-Wallis test shows a significant effect on accuracy from this condition-distance variable, $H(5) = 151.03, p < 0.01$, a large effect ($\eta^2 = 0.343$). To further understand this effect, we perform a pairwise analysis using Wilcoxon rank sum tests with Holm-Bonferroni p-value adjustment, which can be seen in Table I.

F. Why Are Some Sonemes Harder To Comprehend?

These results on the performance of both versions of SIREN at a distance, along with the per-soneme results shown in Figure 8, raise an interesting question: Why are some sonemes harder to comprehend than others, particularly at a distance? One contributing factor to difficulties identifying sonemes at a distance may be their duration. When considering accuracy at distances greater than 1m, a Kruskal-Wallis test indicates a statistically significant effect $H(28) = 78.726, p < 0.01$ with a large effect size ($\eta^2 = 0.889$). Spearman's rank correlation agrees with this finding, showing a statistically significant, positive correlation between soneme length and average soneme accuracy (considering distances $>$ 1m), $r(34) = 0.662, p < 0.01$. This indicates that the longer a soneme is, the easier it is to understand from a distance. This does not fully capture the complexity of soneme design and recognizability, which likely has to do with the frequencies used, their relation to background noise, and the frequency response of the audio production device.

G. Participant Impressions Of SIREN

In the debrief survey, participants were asked to complete a modified version of the Godspeed [32] questionnaire. Participants in their responses indicated positive feelings toward the robot, rating it pleasant ($\mu = 3.83$) and friendly ($\mu = 3.92$). Participants were also given the opportunity to make comments on SIREN. *Participant A* remarked "The higher/more aggressive tones felt easier to hear from a distance than the lower/softer tones...The higher tones seemed to cut through the water white noise *[sic]* much better for me". This reflects the difficulties that participants had identifying sonemes at a distance, but also the fact that some Tonal-Sonemes performed better than others overall.

VII. Discussion And Recommendations

The study presented in this paper evaluated SIREN in a controlled pool environment with swimmers. Applying this system in field environments with divers equipped with scuba gear is likely to involve some amount of adaptation, due to differences in the auditory environment and the added cognitive load of diving and completing a task. First, we recommend expanding the length of all sonemes, by repeating short sonemes a number of times for each requested communication. Section VI-F suggests that longer sonemes are easier to comprehend at a distance, leading to this recommendation. This also allow divers to comprehend the soneme even if portions of the sound are covered by their breathing noises. Additionally, tonal $Attention_S$ sonemes should be used to draw diver attention prior to attempting further communication. This approach would require an algorithm estimating diver attention, either from visual sources or using an acknowledgment model, where the robot continues calling for attention until the diver provides a confirmation signal. While not included in this article due to space constraints, we have completed further research using SIREN in a full-loop communication system, where scuba divers were asked to complete a task using help from the robot. In that task, other forms of communication were frequently missed due to divers focusing on the task, but SIREN often prompted them to pay attention, allowing them to comprehend sonemes or other forms of communication, even while engaged in a task that involved swimming, manipulating small objects, and planning a search pattern. Further exploration is required, but this work indicates that divers should be able to use SIREN in field environments, particularly if the recommendations of this section are applied.

VIII. Conclusion

In this work, we presented SIREN, a device, software system, and two soneme languages for audible communication from AUVs to divers underwater. We first presented the hardware and software design of our system, along with two versions of a sound-based communication language. With our sonemes defined, we performed a human study of soneme perception in underwater environments with 12 participants. The results from this study revealed reasonable accuracies for both forms of sonemes at close distances, with tonal sonemes operating more effectively than text-to-speech sonemes at greater distances. Our analysis of these results also revealed correlations between soneme length and performance at distance, and indicated some possible directions for further improvement of sonemes. SIREN is the first system of its kind for AUVs, providing a low-cost device for AUV-to-diver

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

communication. The results of our work have established a baseline for performance and some strengths and weaknesses of certain types of auditory communication underwater, but many questions still remain in this area, particularly in terms of designing effective communications with sound.

Acknowledgment: We wish to thank the participants and staff of our study, Travis Manderson for suggesting the use of a surface transducer, Brandon Herrera for help in part selection, and Dr. Anna Fulton for her advice on Tonal-Soneme design.

REFERENCES

[1] N. Mišković, M. Bibuli, A. Birk, M. Caccia, M. Egi, K. Grammer, A. Marroni, J. Neasham, A. Pascoal, A. Vasilijević, and others, "CADDY—Cognitive Autonomous Diving Buddy: Two Years of Underwater Human-Robot Interaction," *Marine Technology Society Journal*, vol. 50, no. 4, pp. 54–66, 2016.

[2] M. Fulton, C. Edge, and J. Sattar, "Robot Communication Via Motion: A Study on Modalities for Robot-to-Human Communication in the Field," *ACM Transactions on Human-Robot Interaction*, vol. 11, pp. 15:1–15:40, Feb. 2022.

[3] M. Fulton, M. Mehtaz, J. Sattar, and O. Queeglay, "Underwater Robot-To-Human Communication Via Motion: Implementation and Full-Loop Human Interface Evaluation," in *Robotics: Science and Systems XVIII*, Robotics: Science and Systems Foundation, June 2022.

[4] A. Birk, "A Survey of Underwater Human-Robot Interaction (U-HRI)," *Current Robotics Reports*, Sept. 2022.

[5] R. L. Adams, "Can You Communicate? Underwater?," Apr. 1971.

[6] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and icons: Their structure and common design principles," *Human–Computer Interaction*, vol. 4, no. 1, pp. 11–44, 1989.

[7] S. Yilmazyildiz, R. Read, T. Belpeame, and W. Verhelst, "Review of semantic-free utterances in social human–robot interaction," *International Journal of Human-Computer Interaction*, vol. 32, pp. 63–85, 2016. Place: United Kingdom Publisher: Taylor & Francis.

[8] M. Fulton, A. Prabhu, and J. Sattar, "HREyes: Design, development, and evaluation of a novel method for auvs to communicate information and gaze direction," *To appear at ICRA '23, pre-print available.*, 2022.

[9] Y. Ukai and J. Rekimoto, "Swimoid: Interacting with an underwater buddy robot," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 423–423, Mar. 2013.

[10] B. Verzijlenberg and M. Jenkin, "Swimming with robots: Human robot communication at depth," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4023–4028, Oct. 2010.

[11] M. Fulton, C. Edge, and J. Sattar, "Robot Communication Via Motion: Closing the Underwater Human-Robot Interaction Loop," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 4660–4666, May 2019. ISSN: 2577-087X.

[12] K. J. DeMarco, M. E. West, and A. M. Howard, "Underwater human-robot communication: A case study with human divers," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3738–3743, Oct. 2014.

[13] A. Bonarini, "Communication in Human-Robot Interaction," *Current Robotics Reports*, vol. 1, pp. 279–285, Dec. 2020.

[14] C. Breazeal, "Toward sociable robots," *Robotics and Autonomous Systems*, vol. 42, pp. 167–175, Mar. 2003.

[15] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, Mar. 2003.

[16] H. Chen, X. Liu, D. Yin, and J. Tang, "A Survey on Dialogue Systems: Recent Advances and New Frontiers," *ACM SIGKDD Explorations Newsletter*, vol. 19, pp. 25–35, Nov. 2017. arXiv:1711.01731 [cs].

[17] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, Feb. 1993.

[18] J. Crumpton and C. L. Bethel, "A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech," *International Journal of Social Robotics*, vol. 8, pp. 271–285, Apr. 2016.

[19] F. Eyssel, L. de Ruiter, D. Kuchenbrandt, S. Bobinger, and F. Hegel, "'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 125–126, Mar. 2012. ISSN: 2167-2148.

[20] H. Tennent, D. Moore, M. Jung, and W. Ju, "Good vibrations: How consequential sounds affect perception of robotic arms," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 928–935, Aug. 2017.

[21] D. Moore, H. Tennent, N. Martelaro, and W. Ju, "Making Noise Intentional: A Study of Servo Sound Perception," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, (New York, NY, USA), pp. 12–21, Association for Computing Machinery, Mar. 2017.

[22] F. A. Robinson, M. Velonaki, and O. Bown, "Smooth Operator: Tuning Robot Perception Through Artificial Movement Sound," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21, (New York, NY, USA), pp. 53–62, Association for Computing Machinery, Mar. 2021.

[23] E. Cha, N. T. Fitter, Y. Kim, T. Fong, and M. J. Matarić, "Effects of Robot Sound on Auditory Localization in Human-Robot Collaboration," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, (Chicago, IL, USA), pp. 434–442, Association for Computing Machinery, Feb. 2018.

[24] E.-S. Jee, Y.-J. Jeong, C. H. Kim, and H. Kobayashi, "Sound design for emotion and intention expression of socially interactive robots," *Intelligent Service Robotics*, vol. 3, pp. 199–206, July 2010.

[25] M. Schwenk and K. O. Arras, "R2-D2 Reloaded: A flexible sound synthesis system for sonic human-robot interaction design," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 161–167, Aug. 2014. ISSN: 1944-9437.

[26] "Dayton Audio - DAEX25W-8 Waterproof 25mm Exciter 10W 8 Ohm." https://www.daytonaudio.com/product/1181/daex25w-8-waterproof-25mm-exciter-10w-8-ohm.

[27] E. Nyquist, "eriknyquist/tones," November 2020. Distributed under the Apache 2.0 License, https://github.com/eriknyquist/tones.

[28] hadware, "Voxpopuli," June 2021. Distributed under the MIT License, https://github.com/hadware/voxpopuli.

[29] "eSpeak: Speech Synthesizer." https://espeak.sourceforge.net/.

[30] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken, "The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 3, pp. 1393–1396 vol.3, Oct. 1996.

[31] C. Edge, S. Sakib Enan, M. Fulton, J. Hong, J. Mo, K. Barthelemy, H. Bashaw, B. Kallevig, C. Knutson, K. Orpen, and J. Sattar, "Design and Experiments with LoCO AUV: A Low Cost Open-Source Autonomous Underwater Vehicle," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1761–1768, Oct. 2020. ISSN: 2153-0866.

[32] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics*, vol. 1, pp. 71–81, Jan. 2009.

[33] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Advances in Psychology*, Human Mental Workload, Jan. 1988.

[34] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971. Place: US Publisher: American Psychological Association.

[35] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965. Publisher: [Oxford University Press, Biometrika Trust].

[36] C. E. Spearman, *The proof and measurement of association between two things*. 1901.

[37] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[38] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, pp. 50–60, Mar. 1947. Publisher: Institute of Mathematical Statistics.

[39] M. Tomczak and E. Tomczak, "The need to report effect size estimates revisited. An overview of some recommended measures of effect size," vol. 21, pp. 19–25, Jan. 2014.

[40] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.