A SilentTide: Enabling Silent Speech Communication for Underwater Military Operations

May 7, 2024

Abstract

Underwater communication remains a significant challenge in military operations due to limited bandwidth, high noise levels, and the distortion of traditional audio signals in aquatic environments.To address these obstacles, we introduce **SilentTide**, a novel system that integrates a full-face mask with a bone conduction microphone, bone conduction speaker, air microphone, and high-definition camera. These technologies are leveraged for silent speech recognition, a technique that detects and interprets the movements of a diver's mouth to convey speech without the need for vocalization. By eliminating the exhalation required for speaking, this approach significantly conserves oxygen, a critical resource in underwater environments, while ensuring effective communication. Additionally, SilentTide utilizes audio-visual speech recognition (AVSR) to enhance communication clarity by combining inputs from bone and air conduction microphones with visual data. Advanced signal processing, including super-resolution algorithms, further refines the quality of bone conduction signals to align closely with air conduction fidelity. SilentTide employs sensor fusion and generative AI technologies to reproduce speech signals without vocalization, integrating multimodal inputs to deliver accurate and reliable communication in challenging conditions. Designed to meet the unique requirements of underwater military operations, SilentTide provides an innovative solution to the enduring challenges of underwater communication. By enabling silent speech and integrating advanced signal reconstruction capabilities, this system enhances communication efficiency and operational effectiveness in aquatic environments.

1 Introduction

Effective communication is a cornerstone of modern military operations, yet achieving reliable speech transmission underwater remains a persistent challenge. The unique properties of underwater environments—such as limited bandwidth, high levels of ambient noise, and signal distortion—severely degrade the performance of conventional air conduction (AC) microphones and acoustic communication systems. These limitations impede the clarity and reliability of communication during critical underwater missions, necessitating innovative approaches to overcome these obstacles.

In this work, we introduce a multimodal communication system designed to address these challenges by integrating bone conduction (BC) microphones, silent speech recognition, and audio-visual speech recognition (AVSR). BC microphones capture speech vibrations transmitted through the skull, bypassing the ambient noise and distortions common in underwater environments.[4] However, BC signals are inherently bandwidth-limited and require enhancement to match the fidelity of AC signals.[10] To address this, we employ super-resolution techniques to align BC signals with AC characteristics, ensuring improved speech clarity. [3]

Furthermore, we explore the concept of silent speech recognition, wherein a transmitter-receiver system detects mouth movements to enable speech recognition without vocalization. This approach not only ensures covert communication in scenarios requiring stealth but also extends the applicability of the system beyond traditional acoustic channels. Complementing these modalities, our AVSR framework fuses audio and visual data—including inputs from BC and AC microphones and video streams—enhancing recognition accuracy and robustness in noisy or complex environments.

By combining these technologies, our proposed system provides a robust, stealth-capable solution to the challenges of underwater communication. This work lays the groundwork for advancing military underwater communication systems, offering strategic advantages in both operational efficiency and security.

Keywords include: Bone Conduction Microphone, Ultrasound Probe, Audio-Visual ASR, Super-Resolution Enhancement, Underwater Communication, Silent Speech, Multimodal Integration

2 Literature Review

Underwater communication remains a complex challenge due to factors like bandwidth restrictions, signal attenuation, and noise interference. To overcome these, researchers have explored alternative approaches such as bone conduction (BC) technology, audio-visual speech recognition (AVSR), and silent speech recognition.

BC microphones capture vibrations through the skull, making them resilient to external noise, though their limited bandwidth necessitates enhancement techniques. AVSR, which combines auditory and visual cues, has proven effective in noisy conditions, but its integration with BC microphones in underwater environments remains limited. Silent speech recognition, enabling non-vocal communication by detecting mouth movements, offers potential for covert operations but requires further adaptation for aquatic use.

This section reviews advancements in these areas, focusing on their relevance to underwater applications and identifying gaps that our proposed system aims to address.

3 Preliminaries

This section provides an overview of the foundational technologies and concepts enabling the SilentTide system. Specifically, we discuss the core principles and state-of-the-art techniques in silent speech recognition, bone conduction microphones and speakers, lip reading using a camera, and multimedia integration.[1] These components form the backbone of the SilentTide system, enabling robust and efficient underwater communipcation.

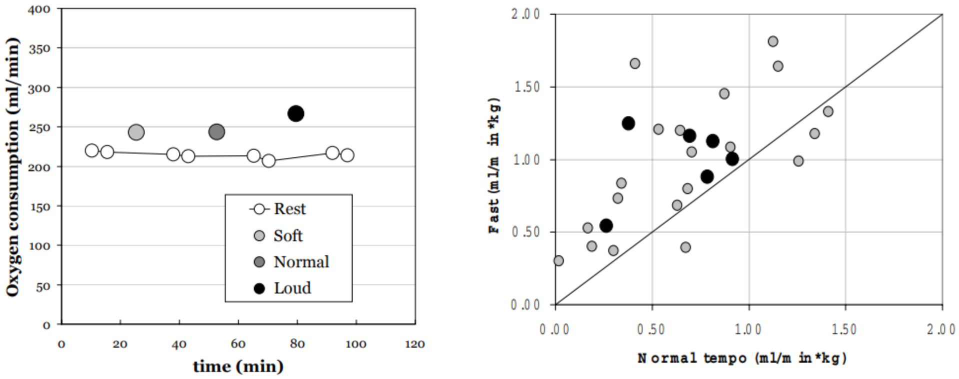
3.1 Oxygen Consumption During Speech Production

Efficient oxygen management is crucial in underwater operations, where divers rely on limited air supplies. Understanding the impact of speech on oxygen consumption is essential for designing communication systems that minimize air usage.

Impact of Vocal Effort and Speaking Tempo

Research by Moon and Lindblom (2003) investigated how vocal effort and speaking tempo affect oxygen consumption during speech production. [5, 8] Their findings include:

- **Vocal Effort**: Oxygen consumption increases with vocal effort. Speaking loudly requires more oxygen than speaking softly.
- **Speaking Tempo**: Faster speaking rates lead to higher oxygen consumption compared to normal speaking rates.



(a) Vocal Effort

(b) Excess O_2

Figure 1: In (a), the figure presents average data for eight subjects from the vocal effort experiment. The ordinate shows the total amount of O_2 used (in ml/min). Open symbols refer to rest conditions. Filled symbols represent speech measurements. In (b), the figure presents excess amounts of O_2 (normalized with respect to subject's body weights) or the normal and the fast rates (in ml/min*kg). A shaded dot pertains to a given session of a specific subject. The black points are subject means. This indicates that subjects used more oxygen during the fast condition.

These results indicate that both the loudness and speed of speech can significantly influence the rate at which oxygen is consumed. In underwater environments, where conserving air is vital, traditional speech can accelerate oxygen depletion due to increased vocal effort and articulatory movements. Implementing silent speech interfaces, such as the SilentTide system, can mitigate these effects by enabling communication without vocalization, thereby preserving the diver's air supply.

By reducing the need for vocal effort and rapid articulatory movements, silent speech systems contribute to more efficient oxygen usage, enhancing the safety and endurance of underwater operations.

3.2 Silent Speech Recognition

Silent Speech Recognition (SSR) refers to the process of interpreting speech without the need for audible vocalization. This is achieved by analyzing articulatory movements, such as lip and tongue positions, or by detecting subvocal signals—neuromuscular activities associated with speech production even in the absence of sound. SSR is particularly beneficial in environments where silence is crucial or where traditional speech is impractical, such as underwater military operations. The performance of SSR systems is typically evaluated using metrics such as Signal-to-Noise Ratio (SNR) and Word Error Rate (WER).[12] SNR measures the clarity of the captured signal relative to background noise, defined as:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \text{dB},$$

where P_{signal} is the power of the desired signal, and P_{noise} is the power of the background noise. A higher SNR indicates a clearer signal, which is crucial for accurate SSR, especially in noisy underwater environments.

WER quantifies the accuracy of the recognized speech by comparing the transcribed output to a reference transcript. It is calculated as:

$$\mathrm{WER} = \frac{S + D + I}{N},$$

where:

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the total number of words in the reference.

A lower WER indicates higher recognition accuracy. In SSR systems, achieving a low WER is challenging due to the absence of acoustic cues, making the integration of multimodal data and advanced processing techniques essential.

By employing SSR within the SilentTide framework, divers can communicate silently, conserving oxygen and maintaining stealth during underwater operations. The combination of visual data analysis and sensor fusion enables effective interpretation of silent speech, addressing the unique challenges of underwater communication.

3.3 Bone Conduction Microphones and Speakers

Bone conduction technology provides an alternative to traditional air-conduction-based audio systems. By transmitting sound through the bones of the skull directly to the inner ear, bone conduction microphones and speakers bypass the challenges posed by ambient noise and other sources of interference.[4]

- **Bone Conduction Microphones**: These devices detect vibrations from the user's speech through contact with the skull. By capturing mechanical vibrations directly, they are less susceptible to ambient noise. For instance, the Motorola Bone Conduction Ear Microphone System utilizes this principle to provide clear communication in noisy environments.
- **Bone Conduction Speakers** These devices convert audio signals into vibrations transmitted through the skull to the inner ear. This method allows users to perceive sound without occluding the ear canal, maintaining environmental awareness—a critical feature for divers.

In underwater scenarios, bone conduction technology offers several key advantages:

- **Noise Resilience**: By bypassing the outer ear, bone conduction devices are less affected by ambient underwater noise, enhancing speech intelligibility.[11]
- **Compatibility with Diving Gear**: Bone conduction microphones and speakers can be integrated into full-face masks without interfering with breathing apparatus or other equipment.
- **Oxygen Conservation**: Enabling communication without the need for vocalization helps conserve oxygen—a vital consideration for divers.

Ultrasound Sensors for Silent Speech Recognition

Ultrasound sensors provide a novel approach to silent speech recognition by capturing articulatory movements through ultrasonic imaging. [2] These sensors utilize high-frequency sound waves to generate detailed images of internal articulatory structures, such as the tongue and jaw, enabling the detection of speech patterns without the need for vocalization. This technology is particularly advantageous in silent environments or scenarios where traditional speech recognition methods are impractical, such as underwater communication.

Principle of Operation

The operation of ultrasound-based silent speech recognition systems involves the following steps:

- **Signal Emission and Capture**: An ultrasonic sensor emits high-frequency sound waves, which interact with the soft tissues of the articulatory system. The reflected waves are captured by the sensor to create a sequence of ultrasound images.

- **Feature Extraction**: These images are processed to extract articulatory movement features corresponding to specific phonemes or words. Key articulatory patterns, such as tongue positions and jaw movements, are identified.

- **Speech Reconstruction**: Advanced deep learning models interpret the extracted features to reconstruct the corresponding speech signal. Techniques such as deep neural networks (DNNs) are employed to map articulatory movements to audio signals with high fidelity.

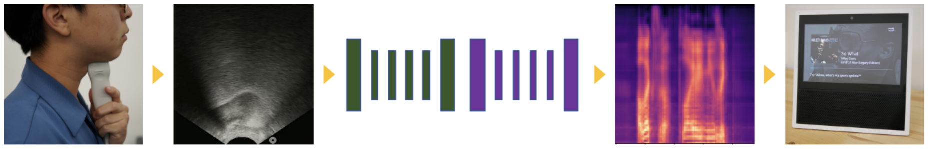


Figure 2: SottoVoce silent voice system: an ultrasonic echo probe attached under the jaw that reads the internal situation while the user is speaking without actually emitting a voice. By recognizing ultrasound images using deep convolutional neural networks, the user's voice is resynthesized and can be used to control the existing speech interaction systems such as smart speakers[2].

3.4 Lip Reading Using Cameras

Lip reading, or visual speech recognition, involves interpreting speech by analyzing visual cues from a speaker's mouth movements. This technique is particularly valuable in environments where audio signals are compromised, such as underwater settings, or in scenarios requiring silent communication.

Traditional Frame-Based Camera Approaches

Conventional lip reading systems utilize frame-based cameras to capture sequences of images depicting a speaker's lip movements. These images are processed to extract features corresponding to phonemes or words. However, frame-based cameras face challenges, including motion blur and limited temporal resolution, which can hinder accurate recognition of rapid articulatory movements.

Advanced methods for isolated single-sound recognition use both frame-based and event-based cameras and generally comprise

- **Imaging from Event Data**: Converting asynchronous event data into images that represent lip movements.

- **Face and Facial Feature Detection**: Identifying and tracking facial landmarks to focus on the lip region.

- **Recognition Using Temporal Convolutional Networks**: Analyzing temporal sequences of lip movements to classify speech sounds.

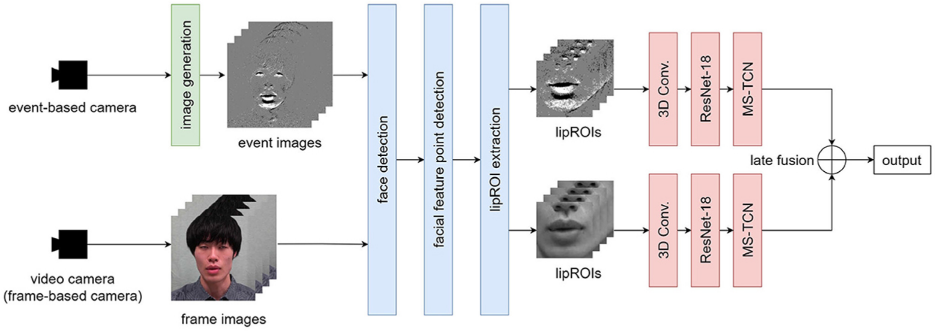


Figure 3: Workflow of a dual-stream lip-reading system using event-based and frame-based cameras from [1]. Event-based cameras capture asynchronous changes, generating event images for feature extraction. Frame-based cameras provide traditional video data. Both streams independently process lip regions (lipROIs) through 3D convolution, ResNet-18, and Multi-Scale Temporal Convolutional Networks (MS-TCN). Outputs are fused in the final stage for accurate speech recognition.

Fig. 5 illustrates the dual-stream architecture used for lip reading of [1]. This approaches leverages both event-based cameras and traditional frame-based video cameras. The top stream represents the workflow of the event-based camera, which captures asynchronous changes in the scene to generate event images. These images undergo face detection, facial feature point detection, and extraction of the lip region of interest (lipROIs). The processed lipROIs are then passed through a 3D convolutional layer, followed by a ResNet-18 feature extractor and a Multi-Scale Temporal Convolutional Network (MS-TCN) for temporal analysis.

The bottom stream follows a similar workflow for the frame-based camera, which captures conventional video frames. The frame images are processed in the same sequence—face detection, facial feature point detection, and lipROI extraction—before being passed through a parallel pipeline of 3D convolution, ResNet-18, and MS-TCN.

The outputs from both streams are fused in the final stage (late fusion) to generate the predicted speech signal, combining the advantages of both camera modalities to enhance lip reading accuracy.

4 System Model: SilentTide Communication Framework

The **SilentTide** system integrates advanced multimodal technologies within a full-face mask to enable robust, oxygen-efficient, and silent underwater communication. The system leverages external and internal components, including bone-conduction microphones and speakers, a high-definition camera, and an ultrasound sensor. These elements collectively capture and reconstruct silent speech for reliable communication in underwater military operations.

A depiction of the proposed SilentTide is provide in Figure 4:

- **External View**: The external configuration of the system showcases the full-face breathing apparatus and visor.

- **Internal View**: The internal view highlights the alignment of the bone conduction speaker and microphone with the user's cranial structure, ensuring effective signal capture and transmission. This perspective also illustrates the integration of the camera system and ultrasound sensor and its interaction with the user's articulatory movements for silent speech recognition.

The system architecture allows for the seamless fusion of inputs from these sensors, which are processed through advanced multimodal sensor fusion algorithms and generative AI models to reconstruct intelligible speech signals for underwater communication.

4.1 Apparatun Construction

This project utilizes the following hardware components to record and process both bone conduction and air conduction signals:

- **Air Microphone:** The air microphone is a traditional microphone that captures air conduction signals by detecting sound waves propagating through the air. This component is crucial for recording speech in environments where bone conduction alone may not provide sufficient detail. The air microphone complements the bone conduction sensor, enabling robust speech recognition and signal reconstruction in multimodal systems.

- **Bone Conduction Sensor:** The Knowles V2S200D is a bone conduction sensor designed to capture vibrations transmitted through the skull. This sensor is ideal for underwater or noisy environments where traditional air conduction microphones may fail to capture clear signals.

- **Arduino Camera:** An Arduino camera will be used to capture visual data, such as lip movements, which will complement the auditory signals from the bone and air conduction microphones. This integration of visual and auditory data will enhance speech recognition and signal reconstruction.

- **Wearable Ultrasound Sensor:** The wearable ultrasound sensor is designed to capture high-frequency signals generated by vocal cord vibrations or other anatomical movements associated with speech. These sensors provide additional data on speech production mechanisms, which can be valuable for silent speech recognition. The sensor will be integrated into the wearable system, ensuring it can operate non-invasively and reliably in real-world scenarios, such as underwater communication or in noisy environments. For this purpose, compact and low-power ultrasound sensors like piezoelectric transducers or MEMS ultrasound sensors are suitable choices.

- **Microcontroller for Signal Recording:** A microcontroller will synchronize the bone conduction and air conduction signals, ensuring the smooth recording and processing of multimodal data. This component will handle the integration of signals for real-time data analysis and reconstruction.

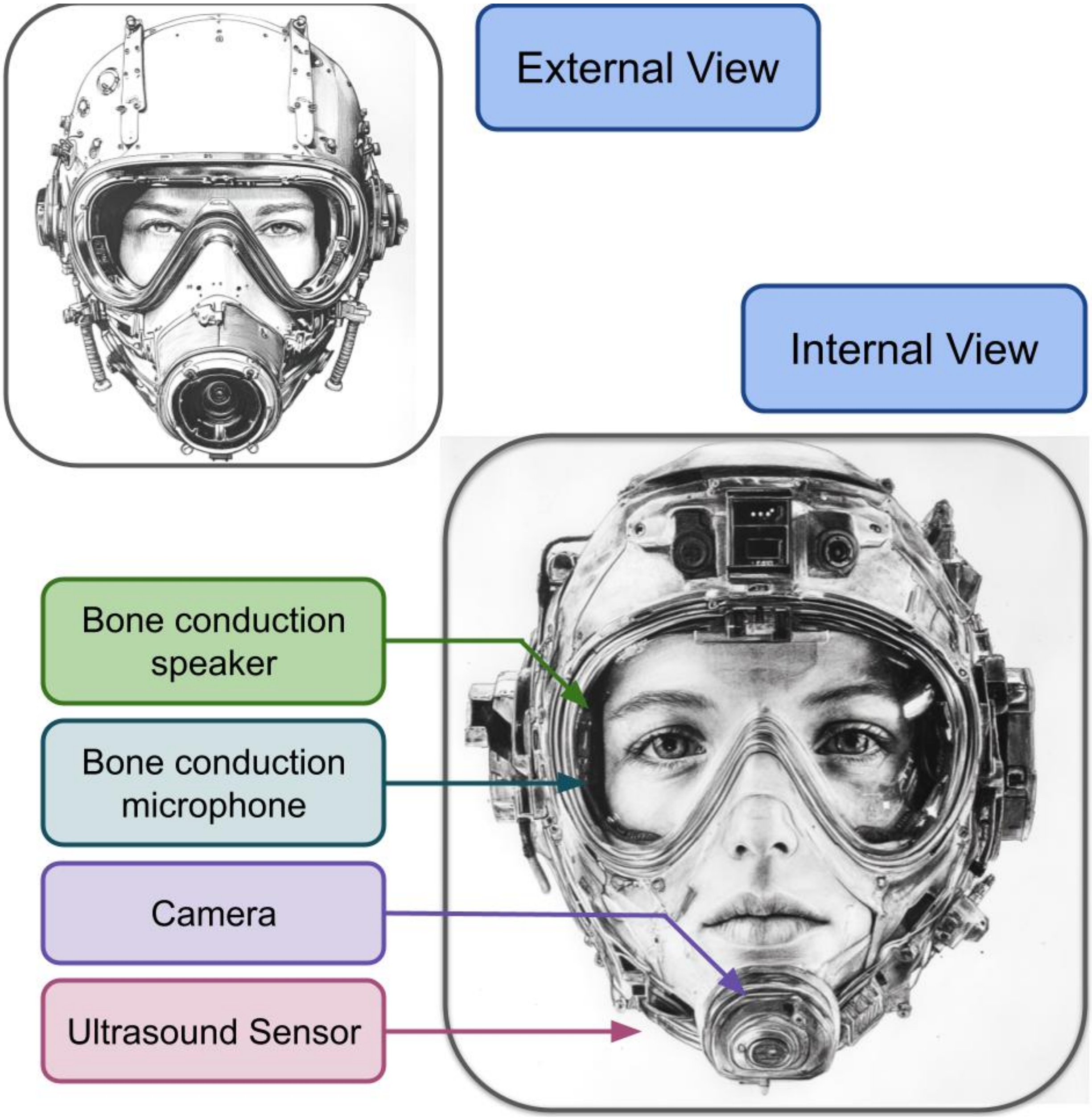


Figure 4: External and Internal Views of the SilentTide System. The external view (top left) highlights the placement of visible components such as the camera and ultrasound sensor, while the internal view (right) focuses on the alignment of bone conduction microphones, speakers, and other components relative to the user's face.

The components will be integrated into the SilentTide system through a series of iterative steps, allowing us to refine their performance and ensure compatibility with the underwater communication framework. Initially, each hardware component will be tested independently to evaluate its performance in capturing and processing silent speech signals. These tests will assess key metrics such as Signal-to-Noise Ratio (SNR) for microphones, image quality for cameras, and sensitivity for the ultrasound sensor.

To facilitate these experiments, we will mount the components on a 3D-printed mask. This prototype setup will allow us to fine-tune the positioning and alignment of the sensors relative to the user's facial structure. The flexibility of a 3D-printed platform ensures that adjustments can be made quickly, enabling iterative optimization of the system's design.

Once the individual components have been validated and optimized, the system will be installed into a commercially available full-face wearable mask. This mask will be adapted to meet the specific requirements of the SilentTide system, including the integration of hardware components and ensuring comfort and usability for the end user. The commercial mask will provide a robust and practical platform for testing the fully integrated system in real-world underwater scenarios.

These stages of experimentation and integration will ensure that the SilentTide system achieves its goals of oxygen-efficient, robust, and silent underwater communication, tailored to the needs of military operations.

4.2 Signal Processing and Multimodal Data Integration

The SilentTide system employs a sophisticated signal processing pipeline to transform raw sensor data into intelligible and actionable speech signals. Following data acquisition from multimodal sources—including bone conduction microphones, air microphones, cameras, and ultrasound sensors—the system processes these signals through a series of advanced techniques.

This section details the key stages of the pipeline, including signal preprocessing, feature extraction, multimodal sensor fusion, and speech reconstruction.

4.3 Multimodal Sensor Sources

For the bone conduction sensor, we are using the V2S200D Knowles, a device designed to selectively capture the speaker's voice while suppressing background noise. This ensures clean and reliable voice recording, making it ideal for underwater or noisy environments where traditional microphones are less effective.

The system employs the following sensor modalities:

- **Bone Conduction Microphone ($x_{\mathbf{BCM}}$):** Captures vibrations transmitted through the skull, isolating the speaker's voice from external noise. This makes it highly effective in challenging environments like underwater or noisy scenarios.

- **Air Microphone ($x_{\mathbf{AM}}$):** Records acoustic signals from the surrounding environment, adding spectral details that complement the BC microphone. This redundancy ensures higher fidelity and captures elements missed by bone conduction.

- **Camera ($x_{\mathbf{CAM}}$):** Provides visual data of lip movements and facial expressions, enabling accurate speech recognition even when audio signals are degraded. This modality enhances performance by incorporating visual speech cues.

- **Ultrasound Sensor ($x_{\mathbf{US}}$):** Detects high-frequency sound waves generated by vocal cord activity, providing supplementary data in environments where other sensors are less effective, such as underwater operations. By employing a flexible and wearable ultrasound device, the system overcomes the limitations of traditional ultrasound recording setups, enabling seamless integration into dynamic and constrained scenarios.

The multimodal input at time t can be represented as:

$$\mathbf{x}(t) = \begin{bmatrix} x_{\text{BCM}}(t) \\ x_{\text{AM}}(t) \\ x_{\text{CAM}}(t) \\ x_{\text{US}}(t) \end{bmatrix},$$

where each modality contributes unique and complementary information for speech recognition.

Integrating these inputs creates a multimodal feature space that compensates for individual sensor limitations, enhancing recognition accuracy across diverse scenarios. Specifically:

- **Noisy environments:** BC microphones filter out external noise, while visual and ultrasound inputs refine recognition when audio signals are degraded.
- **Underwater communication:** Bone conduction and ultrasound sensors provide reliable input where traditional microphones fail.

This multimodal approach ensures robust and resilient performance, meeting the demands of high-stakes applications by maintaining functionality even if one modality is compromised.

4.4 Multimodal Sensor Fusion

Sensor fusion integrates inputs from bone conduction and air conduction microphones and video data to create a robust speech representation. This process is modeled as:

$$\mathbf{z}(t) = f_{\text{fusion}}(\mathbf{x}(t)),$$

where f_{fusion} is a deep learning-based fusion network that aligns and combines temporal and spatial features from each modality.

The fusion network includes:

- **Feature Encoders:** Extract high-level features from audio and video inputs.
- **Cross-Modal Attention:** Dynamically weights informative features across modalities.
- **Temporal Layers:** Aggregate temporal dependencies for coherent speech representation.

This approach enhances noise robustness and compensates for degraded audio using visual cues, ensuring reliable and intelligible speech reconstruction in challenging underwater environments.

4.5 Generative AI for Speech Reconstruction

The fused representation $\mathbf{z}(t)$, which combines features from multimodal inputs such as audio and visual cues, is passed through a generative AI model. This model reconstructs the speech signal with enhanced clarity and robustness. Let $y(t)$ represent the reconstructed speech signal:

$$y(t) = f_{\text{gen}}(\mathbf{z}(t)),$$

where f_{gen} is a generative model, designed to produce intelligible speech signals even in noisy underwater conditions. The architecture of the Audio-Visual Automatic Speech Recognition (AV-ASR) model integrates both auditory and visual modalities to enhance speech recognition performance, particularly in challenging environments. Below is a description of the deep learning components involved:

4.5.1 Audio Stream Processing

The audio input, processed from both bone conduction (BC) and air conduction (AC) microphones, is passed through a **Convolutional Neural Network (CNN)** or **Recurrent Neural Network (RNN)** to extract acoustic features. These networks are capable of handling the sequential and temporal dependencies in speech data, learning to capture essential audio patterns, such as phonemes, intonations, and speech rhythm.

4.5.2 Visual Stream Processing

Simultaneously, visual input (e.g., lip movements) is captured by the Arduino camera and processed by a **CNN**. This network extracts facial features that are crucial for visual speech recognition. Facial features provide supplementary information that helps improve the accuracy of speech recognition, especially in noisy or underwater environments where audio quality may degrade.

4.5.3 Multimodal Fusion

The extracted audio and visual features are fused using an **attention mechanism** or a **multimodal fusion layer**. This fusion layer combines the temporal and spatial information from both modalities, emphasizing the most relevant features at each time step. Fusion allows the model to benefit from both the auditory and visual cues, improving robustness in environments with high noise levels.[6, 7]

4.5.4 Voice Activity Detection (VAD)

A secondary branch is dedicated to **Voice Activity Detection (VAD)**, which helps identify active speech periods. This VAD task ensures that the system can ignore silent or non-speech segments, improving efficiency and reducing computational load.

4.5.5 Automatic Speech Recognition (ASR)

The primary task of the model is **ASR**, which processes the fused audio-visual features to transcribe speech into text. This task is tackled using **transformer networks** or **fully connected layers**, which learn to map the fused multimodal data to the target transcription.

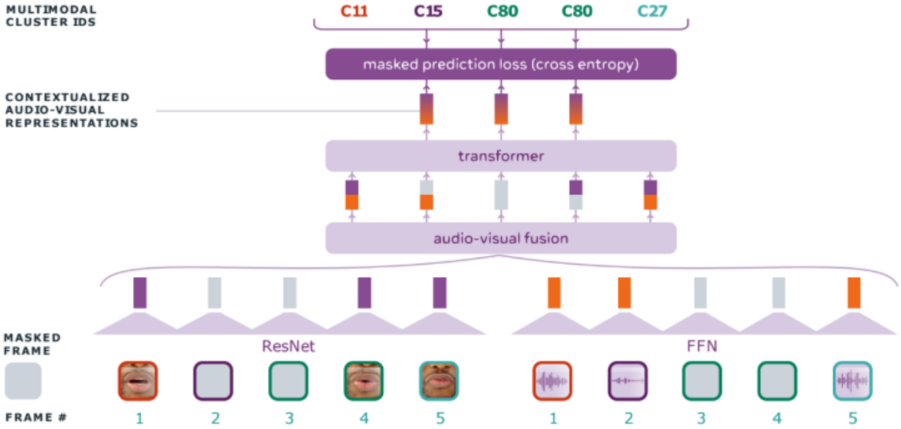


Figure 5: AV-HuBERT, as shown in the figure, is a model that combines audio features with the visual features. More formally, given an audio stream $A = [a_1, \ldots, a_R]$ and a visual stream $I = [i_1, \ldots, i_T]$ aligned together[9].

Hybrid Approaches A hybrid model combining VAEs and GANs could also be implemented to leverage the strengths of both architectures. The VAE ensures stability and generalization, while the GAN enhances realism and clarity. This dual approach is particularly valuable for underwater communication, where signal integrity and naturalness are critical.

The use of these generative AI models represents a novel approach to speech reconstruction, addressing the challenges of underwater environments by ensuring clarity, robustness, and intelligibility.

4.6 Underwater Ultrasound Communication System

The reconstructed speech signal $y(t)$ is converted into an ultrasound-compatible signal for transmission:

$$s_{\text{ultrasound}}(t) = f_{\text{ultrasound}}(y(t)),$$

where $f_{\text{ultrasound}}$ is a modulation function that adapts the signal for underwater communication.

This framework ensures robust, oxygen-efficient communication by leveraging multimodal inputs, advanced sensor fusion techniques, and generative AI for seamless underwater speech transmission.

5 Workplan

The project is structured into a series of work packages (WP), each focusing on a critical stage of development. The timeline considers the complexity of each task, allocating appropriate durations to ensure the systematic completion of the project.

Work Package 1: Data Recording with Mandarin Speech

Duration: 3 months
Justification:

- **Set up sensors and microphones (1 month):** Connect and synchronize the Knowles V2S200D bone conduction sensor, air conduction microphones, and Arduino camera with the microcontroller. Ensure reliable operation and data synchronization.
- **Data collection (1 month):** Record Mandarin speech in diverse conditions, including noisy, clean, and underwater environments, to capture varied acoustic features.
- **Preprocessing (1 month):** Clean and preprocess data, ensuring it is segmented and normalized for consistent quality across the training, validation, and test sets.

Work Package 2: Sensor Fusion

Duration: 4 months
Justification: Fusion of audio and visual data requires alignment of signals and designing robust multimodal integration techniques.

- **Synchronize sensors (1 month):** Ensure temporal alignment of BC, AC, and video signals for consistent multimodal data streams.
- **Data fusion algorithm (2 months):** Develop and train a fusion model, incorporating attention mechanisms or multimodal layers to combine audio and visual modalities effectively.
- **Evaluation (1 month):** Validate the fused representation to ensure robustness across different environments and use cases.

Work Package 3: Training a Model for Reconstructing BC to AC

Duration: 6 months
Justification: Developing a model to reconstruct AC signals from BC inputs involves significant experimentation and computational effort.

- **Model architecture (1 month):** Design a model (e.g., VAE or GAN) to learn the mapping from BC signals to AC signals.
- **Loss function and training (4 months):** Define reconstruction losses and train the model using diverse datasets to ensure speaker- and environment-agnostic performance.

- **Validation (1 month):** Test the model on unseen data to verify generalization and fidelity in AC reconstruction.

Work Package 4: Training AV-ASR with Extra BC Signal (Multitask Model)

Duration: 7 months
Justification: Designing a multitask model for ASR and VAD, incorporating BC signals, demands careful architecture and extensive training.

- **Model design (2 months):** Develop a multitask model that integrates multimodal data and optimizes for ASR as the primary task and VAD as a secondary task.
- **Data augmentation and training (4 months):** Augment data with noisy and clean speech datasets, training the model for robustness in diverse environments.
- **Evaluation (1 month):** Ensure a balanced loss for ASR and VAD tasks, validating the model's effectiveness.

Work Package 5: Inference and Productization

Duration: 4 months
Justification: Integrating models into a deployable system requires engineering but is less research-intensive than training phases.

- **Inference pipeline (1 month):** Develop a pipeline for real-time inference, integrating BC-to-AC reconstruction and AV-ASR models.
- **System integration (1 month):** Combine the models into a unified product with seamless functionality.
- **Optimization (1 month):** Ensure low latency and high accuracy for real-time operation.
- **Product packaging (1 month):** Package the system into a user-friendly wearable or mobile application.

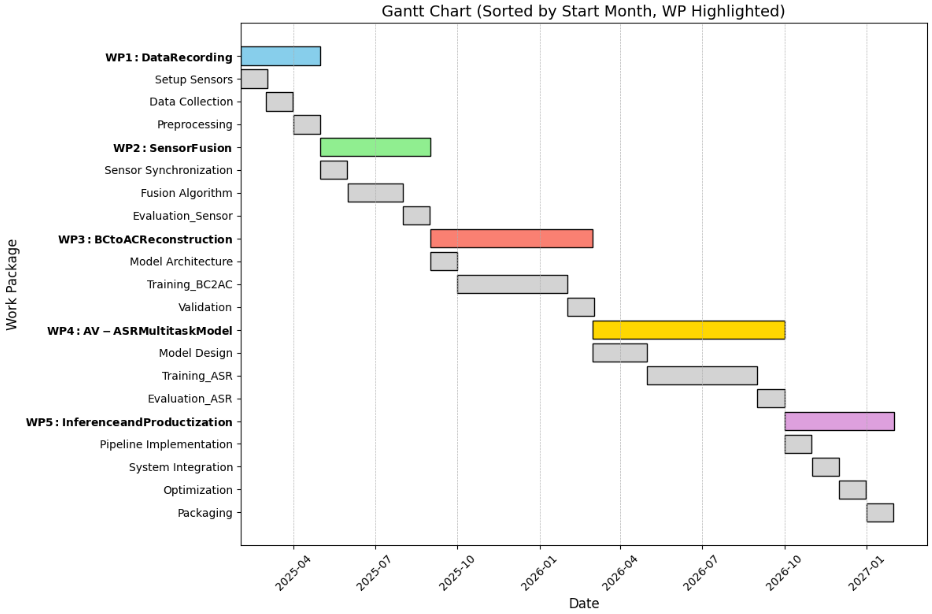


Figure 6: Work Plan with specific duration

References

[1] Tatsuya Kanamaru, Taiki Arakane, and Takeshi Saitoh. Isolated single sound lip-reading using a frame-based camera and event-based camera. *Frontiers in Artificial Intelligence*, 5:1070964, 2023.

[2] Naoki Kimura, Michinari Kono, and Jun Rekimoto. Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[3] Yuang Li, Yuntao Wang, Xin Liu, Yuanchun Shi, Shwetak Patel, and Shao-Fu Shih. Enabling real-time on-chip audio super resolution for bone-conduction microphones. *Sensors*, 23(1):35, 2022.

[4] Maranda McBride, Phuong Tran, Tomasz Letowski, and Rafael Patrick. The effect of bone conduction microphone locations on speech intelligibility and sound quality. *Applied ergonomics*, 42(3):495–502, 2011.

[5] Seung-Jae Moon and Björn Lindblom. Two experiments on oxygen consumption during speech production: vocal effort and speaking tempo. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 3129–3132, 2003.

[6] R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L Koerich, Simon Bacon, Patrick Cardinal, et al. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2486–2495, 2022.

[7] R Gnana Praveen, Eric Granger, and Patrick Cardinal. Cross attentional audio-visual fusion for dimensional emotion recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[8] Allison L Rosenthal, Soren Y Lowell, and Raymond H Colton. Aerodynamic and acoustic features of vocal effort. *Journal of Voice*, 28(2):144–153, 2014.

[9] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022.

[10] Ho Seon Shin, Hong-Goo Kang, and Tim Fingscheidt. Survey of speech enhancement supported by a bone conduction microphone. In *Speech Communication; 10. ITG Symposium*, pages 1–4. VDE, 2012.

[11] Putta Venkata Subbaiah and Hima Deepthi. A study to analyze enhancement techniques on sound quality for bone conduction and air conduction speech processing. *Scalable Computing: Practice and Experience*, 21(1):57–62, 2020.

[12] Asma Trabelsi, Sébastien Warichet, Yassine Aajaoun, and Séverine Soussilane. Evaluation of the efficiency of state-of-the-art speech recognition engines. *Procedia Computer Science*, 207:2242–2252, 2022.