Human Detection And Tracking For Human-Robot Interaction On The REEM-C Humanoid Robot

by

Pranav Barot

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis contains four conference manuscripts prepared throughout the master's research program.

1. Pranav Barot, Ewen MacDonald, Katja Mombaur, Vision Systems For Identifying Interlocutor Behaviour And Augmenting Human Robot Interaction, In Conference on Vision and Intelligent Systems (**CVIS**), Waterloo, Canada, 2022. **Winner of the Best Imaging Paper Award**.
2. Pranav Barot, Ewen MacDonald, Katja Mombaur, Natural Head And Body Orientation For Humanoid Robots During Conversations With Moving Human Partners Through Motion Capture Analysis, In Conference on Advanced Robotics and its Social Impact (**ARSO**), Berlin, Germany, 2023. **Finalist for the Best Paper Award**.
3. Pranav Barot, Ewen MacDonald, Katja Mombaur, An Audio-Video Sensor Fusion Framework To Augment Humanoid Capabilities For Identifying And Interacting With Human Conversational Partners, In International Conference on Humanoid Robots, Austin, USA, 2023 **(UNDER REVIEW)**
4. Pranav Barot, Katja Mombaur, Ewen MacDonald, Estimating Speaker Direction On A Humanoid Robot With Binaural Acoustic Signals, In Journal Of Public Library Of Science (PLOS ONE), 2023 **(UNDER REVIEW)**

The content of these papers is used directly in this thesis, with some adaptation and further explanation where necessary. The papers are organized into individual sections and in a manner that best describes their contribution to the overall thesis.

Abstract

The interactions between humanoid robots and humans is a growing area of research, as frameworks and models are being continuously developed to improving the ways in which humanoids may integrate into society. These humanoids often require intelligence beyond what they are originally endowed with in order to handle more complex human-robot interaction scenarios. This intelligence can come from the use of additional sensors, including microphones and cameras, which can allow the robot to better perceive its environment. This thesis explores the scenarios of moving conversational partners, and the ways in which the REEM-C Humanoid Robot may interact with them. The additional developed intelligence focuses on external microphones deployed to the robot, with a consideration for computer vision algorithms built using the camera in the REEM-C's head.

The first topic of this thesis explores how binaural acoustic intelligence can be used to estimate the direction of arrival of human speech on the REEM-C Humanoid. This includes the development of audio signal processing techniques, their optimization, and their deployment for real-time use on the REEM-C.

The second topic highlights the computer vision approaches that can be used for a robotic system that may allow better human-robot interaction. This section describes the relevant algorithms and their development, in a way that is efficient and accurate for real-time robot usage.

The third topic explores the natural behaviours of humans in conversation with moving interlocutors. This is measured via a motion capture study and modeled with mathematical formulations, which are then used on the REEM-C Humanoid Robot. The REEM-C uses this tracking model to follow detected human speakers using the intelligence outlined in previous sections.

The final topic focuses on how the acoustic intelligence, vision algorithms and tracking model can be used in tandem for human-robot interaction with potentially multiple human subjects. This includes sensor fusion approaches that help correct for limitations in the audio and video algorithms, synchronization and evaluation of behaviour in the form of a short user study. Applications of this framework are discussed, and relevant quantitative and qualitative results are presented.

A chapter to introduce the work done to establish a chatbot conversational system is also included.

The final thesis work is an amalgamation of the above topics, and presents a complete and robust human-robot interaction framework with the REEM-C based on tracking moving conversational partners with audio and video intelligence.

Acknowledgements

I would like to thank all the people who made this thesis work possible.

Dr. Ewen MacDonald, for his guidance and support throughout my degree, in the context of designing and testing the signal processing algorithms and how they may be used in real-time for the REEM-C.

Dr. Katja Mombaur, for the opportunity to work with the HCRMI Lab and for her feedback and guidance on developing and integrating real-time HRI systems for the REEM-C.

I would also like to extend thanks to Dr. Yue Hu and Dr. Paul Fieguth for their valuable efforts in reading and reviewing this thesis work.

I would also like to thank all members of the HCRMI Lab for their support during my degree, whether it be through technical advice, helping evaluate the developed robotics systems or filming demo videos.

Finally, I would like to thank my parents for the support provided throughout my education, and the encouragement to pursue an MASc at the University of Waterloo.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Robotic systems, more specifically humanoids, are becoming more and more integral to a society that is depending on automation to assist humans in their everyday tasks. This automation applies to a variety of areas, including robotics for manufacturing and assembly tasks, social robotics, security and surveillance, and more. These environments require humans to perceive their environment and work alongside other humans through both verbal and non-verbal communication. In order for humanoid robots to effectively handle these complex scenarios, they will require intelligence beyond what they have been originally endowed with in order to perceive and respond in a similar manner to humans.

In the context of human-robot communication, an acoustic form of intelligence may be required to respond to auditory stimuli and maintain conversational dynamics. Human-robot workspaces will involve humans verbally communicating with each other to provide instructions and feedback, amongst other auditory stimuli, such as tools being used or tables being moved around. Humans will also be moving around in these environments, and so humanoid robots will require the ability to communicate with human subjects and respond to this motion in a natural and intelligent way. The usage of acoustic intelligence on a humanoid robot is therefore of research interest, more specifically, identifying the direction of arrival of human speech and being able to reject other non-speech sounds. Acoustic direction of arrival techniques have been studied [30], but their integration with robotics systems in real-time, and the ways in which the robotic systems should respond, is still a challenging problem.

The limitations of acoustic intelligence may still be ameliorated with the usage of vision. Humanoids often have the capability to use vision systems, which may complement the acoustic intelligence and vice versa. For use in real-time, these vision systems will have to be lightweight, fast, and designed to adhere to the real visual data that the robot will perceive as opposed to external datasets that may not be representative (in terms of resolution and frame rate, for instance). This thesis also explores relevant vision systems for human-robot interaction. Sensor fusion techniques that use both the vision and audio information in tandem are also designed and evaluated.

The desired behaviour of humanoids in these scenarios is another topic of interest. Past work such as [15] explores how humans use their body to turn towards targets during locomotion. With the potential to detect the angular displacement of human subjects, a humanoid will also need to naturally reorient to the moving interlocutor. This behaviour will need to be realistic and human-like to allow for humans to accept their presence in everyday scenarios. A method to measure, model and evaluate this behaviour is also presented in this thesis. This behaviour is then integrated with the previously described systems, resulting in a more complete and robust HRI framework.

The overall goal of this work is to explore a variety of techniques with which the REEM-C Humanoid Robot can be made more aware of its environment in a human-robot interaction context, and how the robot should respond given a variety of scenarios. This work has direct applications to social robotics and human-robot collaboration, and distinguishes itself by integrating a number of different intelligent systems to augment the capabilities of the REEM-C humanoid robot.

1.1 The REEM-C Humanoid Robot

The robotic system used for this work is the REEM-C Humanoid, manufactured by PAL Robotics in Spain. It has a combined 68 degrees of freedom, stands at a height of 165cm, and has a weight of 80kg. Fig 1.1 shows the REEM-C used for this work at the University of Waterloo.

The human-like appearance makes this robot ideal for developing human-robot interaction frameworks. This work adds to the capabilities of the REEM-C by installing microphones on the head to facilitate binaural audio processing. This installation is done in a way that replicates the configuration of human ears, and maintains a realistic, natural appearance for the REEM-C, as opposed to other potential solutions such as the ReSpeaker or the Amazon Alexa.

This thesis places a direct focus on the abilities of the REEM-C to reorient itself using the head, torso and feet. The head has two degrees of freedom, which control its yaw and pitch. The torso of the REEM-C has the same two degrees of freedom, which control its

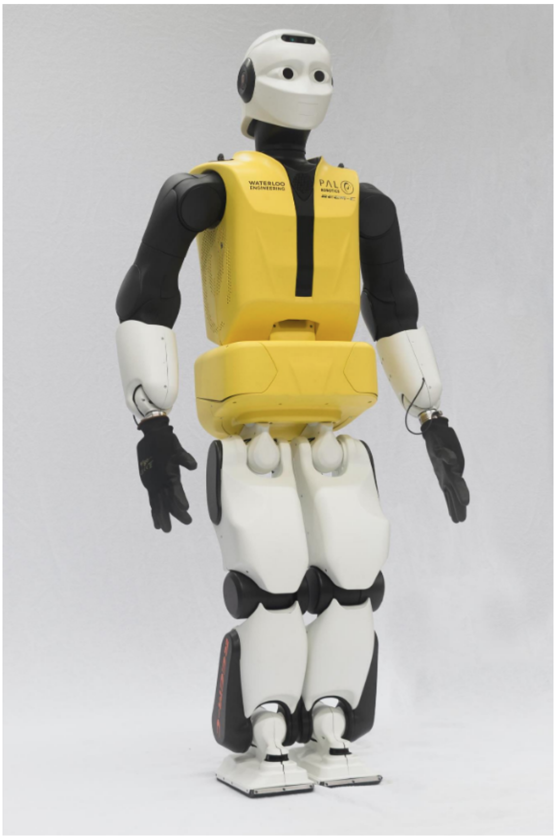

Figure 1.1: REEM-C

yaw and pitch allowing for more expressive movements. The legs, with 6 DoF each, are controlled via a stepping behaviour, implemented via a ROS client provided by PAL, which controls the size and orientation of the steps made by both legs.

The specific usage of these DoF are outlined in Chapter 4. Figures 1.2 and 1.3 show the upper and lower kinematic chains of the REEM-C. Each coordinate frame is located at the joint, and each tree structure begins at the base link, which is the pelvis, where the center of mass of the REEM-C lies.

Hence, for the upper body, the joints of interest in this work are *head 2* and *torso 2* as they control the yaw of the REEM-C. For the lower body, stepping is controlled via a higher level interface that resolves step size and change in orientation to the required joint angles. The details for how these joints are controlled are explained in Chapter 4.

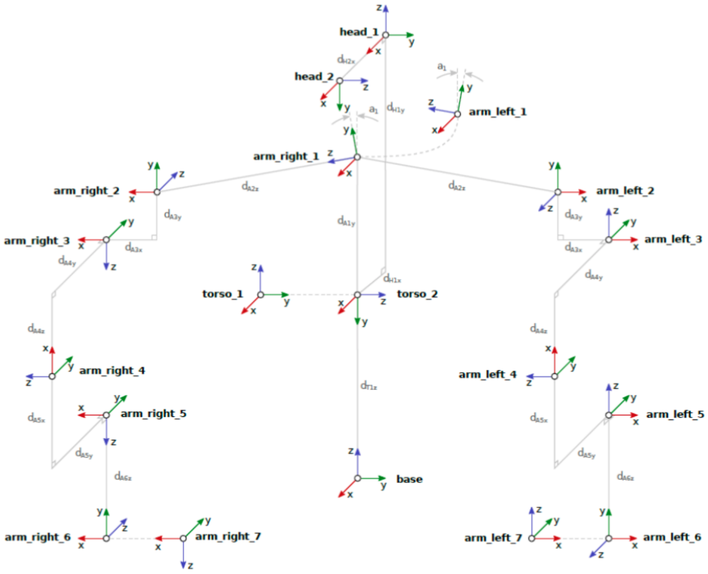

Figure 1.2: REEM-C upper-body kinematic chain

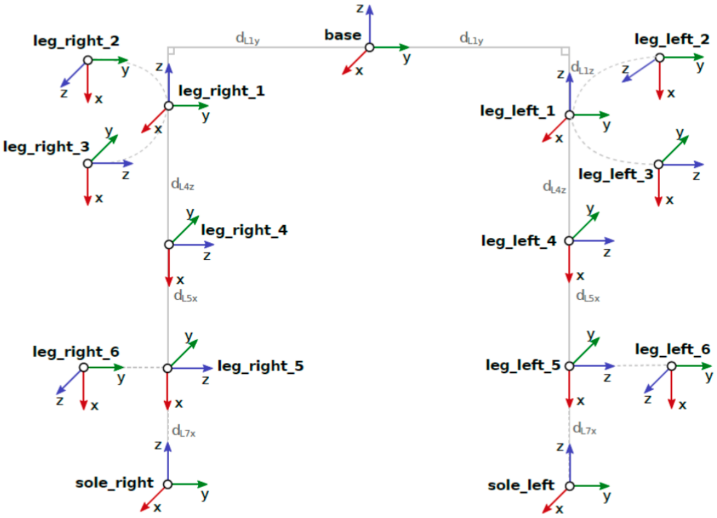

Figure 1.3: REEM-C lower-body kinematic chain

1.2 Problem Definition

This thesis approaches human-robot interaction from multiple perspectives.

The first is the development of sound source localization capabilities for the REEM-C Humanoid Robot, which can be used for detecting human speech and the direction from which the speech arrives. This can be done with a binaural setup, applying real-time audio digital signal processing concepts.

The second perspective is to design algorithms to identify the behaviours of human interlocutors using the vision capabilities of the REEM-C. These behaviours are focused on the minutiae of human behaviours, such as gaze estimation and voice activity detection with visual cues. The visual direction of arrival is also measured, as the supplement to the acoustic direction of arrival.

When it comes to moving human subjects, it is necessary to develop a framework with which the REEM-C may orient itself and track the interlocutor. This model can be designed from data collected via motion capture studies, allowing for the REEM-C to exhibit natural human-like behaviours modeled after real human interactions.

Finally, there is a need to effectively integrate all developed systems together at once for operation on the REEM-C. This involves sensor fusion techniques for the audio and video, as well as methods to account for the shortcomings of each sensor. The potential of conversational agents to allow for a verbal interaction between human subject and robot is also explored from a technical integration point of view.

1.3 Thesis Organization

The organization of this thesis conforms to the key points presented in the problem definition.

Chapter 2 outlines the audio signal processing pipeline and optimization used for real-time deployment on the REEM-C.

Chapter 3 describes the relevant vision systems for the human-robot interaction framework.

Chapter 4 reports on the motion capture study and modelling performed to allow for the REEM-C to track moving conversational partners in a natural way.

Chapter 5 describes the sensor fusion techniques and integration for the audio and video intelligence, which improves the capabilities of the REEM-C to interact with multiple subjects at once.

Chapter 6 briefly outlines initial work done to add a conversational pipeline that utilizes real-time audio streaming and response generation.

Finally, chapter 7 concludes with an overview of the thesis work and considerations for the future.

Chapter 2

Audio Signal Processing For Direction of Arrival Estimation

2.1 Introduction

Speech is one of the most important forms of human communication and a key element of social interaction. Thus, to better integrate humanoid robots into society and augment human-robot interaction, it is important for them to achieve speech interactions that are similar to human-human interactions. Speech interactions are a complex phenomenon that includes both verbal and non-verbal behaviour. One aspect of this non-verbal behaviour is how talkers and listeners orient their head and body relative to their conversational partner.

In the present study, we focus here on a sub-task of identifying the direction of arrival (DOA) of human speech. This is information is necessary for humanoid robots to interact with humans in realistic and natural ways, such as orienting to and tracking human conversational partners (who may move during the conversations), or handling interactions that involve multiple conversational partners. [1]

Much work has been done on sound source localization (SSL) by robots (for a review see [30]) and many of the methods are based on cues that are used by humans to localize sound sources. Given an array of two or more microphones that are spatially separated,

[1] The content of this chapter is from the following journal paper: Pranav Barot, Katja Mombaur, Ewen MacDonald, Estimating Speaker Direction On A Humanoid Robot With Binaural Acoustic Signals, In Public Library On Science (PLOS ONE), 2023

the sound from a source will arrive at each microphone at different times. Thus, by measuring the time difference of arrival between microphones, and knowing the geometry of the microphone array, it is possible to estimate the DOA of the source. This method is analogous to the use of inter-aural timing (ITD) difference cues used by humans. A related approach involves the use of beamforming. The output level of a beamformer should be higher if it is steered in the direction of the source. Thus, DOAs can be estimated by finding look directions which correspond to maxima of the beamformer output levels. If an object is present between the microphones in an array, that object will alter the acoustic field and can vary the level of the signals received at the different microphones. For example, if the object is large compared to the wavelength of the source, the object can cast an acoustic "shadow". Thus, microphones where the object is located in the direct path to the source will record lower levels than those where the object is not in the path. This is analagous to inter-aural intensity differences (IID) used by humans (where the head can result in substantial level differences between the ears at high frequencies). For different DOAs, the geometry of the irregularly-shaped human pinnae (the part of the ear that is on the head) results in patterns of constructive and destructive interference that will vary with DOA. These spectral notches "colour" the sound received by the ear. Thus, by estimating the patterns of spectral notches, it is possible to infer the DOA. Given the complexity of these patterns and the relationship with DOA, this used of this spectral approach relies on learning methods.

A further factor to consider in SSL is the effect of the environment. In general, sound sources radiate sounds in multiple directions. Surfaces that are present in the environment (e.g., walls, floor, ceiling, furniture, etc.) will reflect a portion of the incident sound. Thus, the sound signal recorded at a microphone will be a sum of the acoustic signal from the direct path between the source and the microphone and all the other paths that involve one or more reflections. In the context of DOA estimation, the paths that involve reflection will have a different DOA than that of the direct path.

In the context of human speech interactions, another key factor is the timing of turns. Previous work investigation human conversation has found that talkers start their turn approximately 200-300 ms after their partner has finished their turn[9, 45, 23]. To achieve human-like interactions, it is necessary for a humanoid robot to respond within a similar time frame. The latency of generating DOA estimates will limit how quickly a humanoid robot can respond to movement of a current talker or orient towards a new talker. Works such as [24, 40] consider accurate DOA estimation on robotic systems, but also require a consideration for latency and turn-taking in the context of human-robot conversational scenarios.

Our work evaluates and optimizes a pipeline consisting of two main stages. The first stage continuously generates DOA estimates based on the acoustic signals received from two microphones placed on the head of a humanoid robot. The second stage categorizes these DOA estimates as being "good", that is the estimate likely corresponds with the direct signal from a human talker rather than background noise or a reverberant echo. Using a manually collected and labeled dataset, we investigate the performance of the pipeline's ability to detect direct human speech among background noise and self-generated robot sounds, accurately estimate the direction of arrival, and account for latency of detection. The unique parameters of the pipeline are optimized via either a brute force approach or a more efficient and useful Bayesian optimization approach, which sheds light on how the pipeline's performance depends on each chosen parameter.

2.1.1 DOA Estimation

The first main stage in the pipeline is to generate DOA estimates based on the acoustic signals received at multiple microphones. In the present study we consider the case where there are two spatially separated microphones. Using two microphones as the robot's "ears" is preferable to more complex arrays to minimize the associated computational expenses, and also provides a more human-like appearance for the robot. For this case, the simplest approach to estimate direction of arrival is to examine the cross-correlation of signals from the two microphones to estimate the difference in arrival time between the two microphones. These received signals can be streamed in real-time, or can be processed after being recorded. The estimate of difference in arrival time can then be resolved to a direction given that the geometric setup of the microphones is known.

Cross-Correlation and Beamforming

Beamforming is a method used to improve the directionality of an array of receivers. A simple method is a delay-and-sum technique, where the signals from each receiver are delayed by a fixed amount that varies across receivers and are then summed together. In this way, the direction of the beam (i.e., direction in which the response from the spatial filter is largest) can be steered by varying the delays. As noted earlier, one can estimate a DOA using a beamformer by steering the beam across all angles and finding the direction that results in the largest signal. When the array consists of only two microphones, the delay-and-sum beamforming technique is closely related to the cross-correlation based DOA method to estimate the maximal time alignment/beam direction.

Consider two waves measured at receiver 1 and receiver 2, as per the equations below, with some added Gaussian noise.

$$\begin{aligned} y_1 &= 2sin(x-1) + \mathcal{N}(0.5,\, 1.2) \\ y_2 &= 3cos(x-0.5) + \mathcal{N}(0.5,\, 1.2) \end{aligned} \tag{2.1}$$

Their raw measured amplitude over time appears as in Fig 2.1.

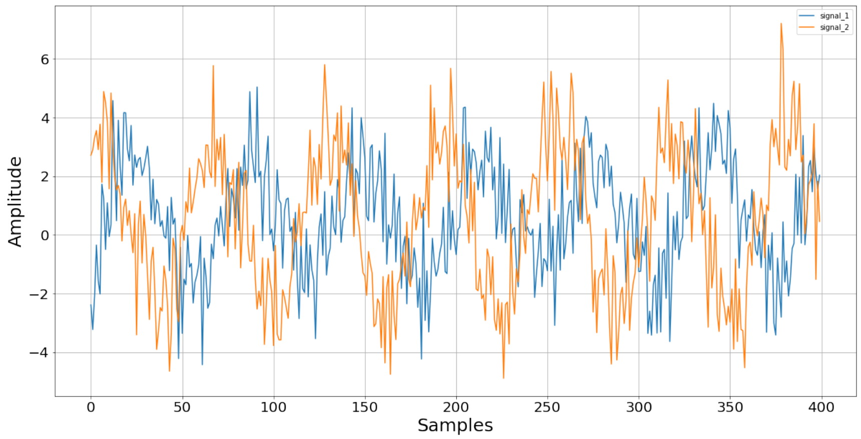

Figure 2.1: Two Separate Waves Visualized

Applying a time domain cross-correlation operation directly results in an output as in Fig 2.2.

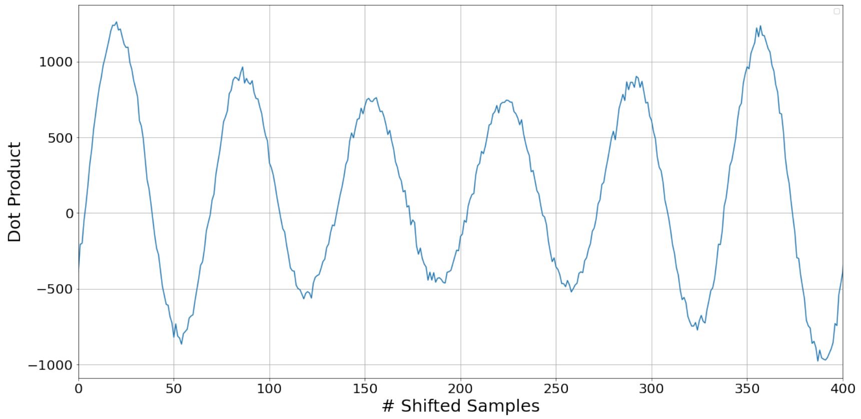

Figure 2.2: Cross Correlation Output of Given Waves

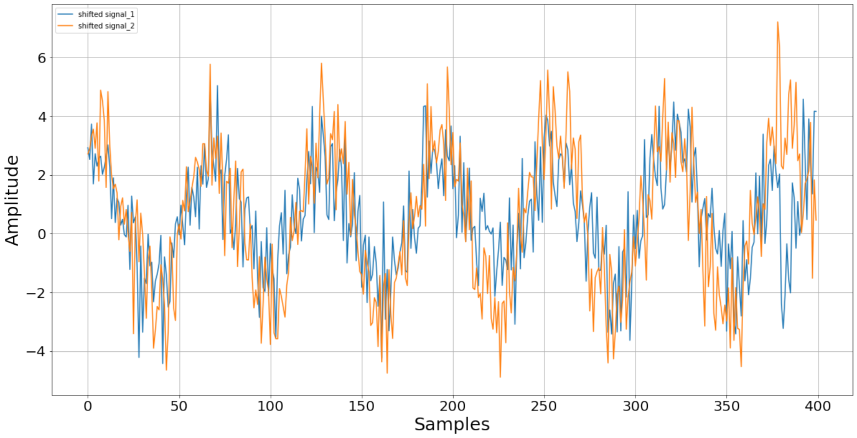

Figure 2.3: Two Waves Aligned After Shift of N=20

The maximum value appears at n=20 samples, indicating that this value best aligns the two received signals. After shifting one signal by the required 20 samples, the resultant is now as in Fig 2.3. Evidently, the signals are well aligned after using the estimate from the cross-correlator.

Variations On Cross Correlation

Since traditional cross-correlators are computationally expensive and sensitive to background noise and reverberation, spectral domain methods are used in this work. Interaural timing differences are estimated using the Wiener-Khinchin relation for the cross-power spectrum, using the Fourier transforms of two recorded signals x and y.

$$G_{xy} = X[f]Y[f]^* \tag{2.2}$$

This relation is used to estimate cross-correlation output of x and y as per the following generalized formulation, the *argmax* of which indicates the ITD between the two microphones [20].

$$\hat{R}_{xy} = \int_{-\infty}^{\infty} \psi(f) G_{xy}(f) e^{j2\pi f\tau} df \tag{2.3}$$

The cross-correlation vector is then the inverse Fourier transform of this result.

Spectral Domain Cross Correlation

The spectral domain cross correlation comes with no whitening transform on the cross correlator. This results in the general estimator as in Eq (2.4). The advantage is the computational efficiency of not requiring a delay-and-sum operation in the time domain while still generating an estimate of the cross-correlation output.

$$\psi_{CC}(f) = 1 \tag{2.4}$$

Generalized Cross Correlation - Phase Transform

The phase transform (GCC-PHAT) pre-whitens the cross-correlation response using the value of ψ as in Eq (2.5), providing robustness against reflections in difficult auditory environments [30].

$$\psi_{PHAT}[f] = \frac{1}{|G_{xy}(f)|} \tag{2.5}$$

Generalized Cross Correlation - Smoothed Coherence Transform

The smoothed coherence transform (SCOT) aims to reduce the error contributed by both signals X and Y, where the PHAT may not be able to adequately handle the case where $G_{xx} \sim 0$ or $G_{yy} \sim 0$ in lower frequency bands. This provides the SCOT pre-whitening, as in Eq (2.6).

$$\psi_{SCOT}[f] = \frac{1}{\sqrt{G_{xx}(f)G_{yy}(f)}} \tag{2.6}$$

These cross-correlation methods are visualized by their output on a frame of 350ms containing speech. Fig 2.4 shows the results from a time domain cross correlation, a frequency domain cross correlation, and the GCC-PHAT.

The two naive cross-correlation methods generate noisier outputs, as their local maxima are quite similar to the global maxima. This is attributed to reflections and reverberation present in the audio frame, which make this problem more complex. However, the GCC-PHAT is able to find one peak that is far more prominent than the rest, as a result of the applied pre-whitening transform. The prominence of the peak increases the confidence

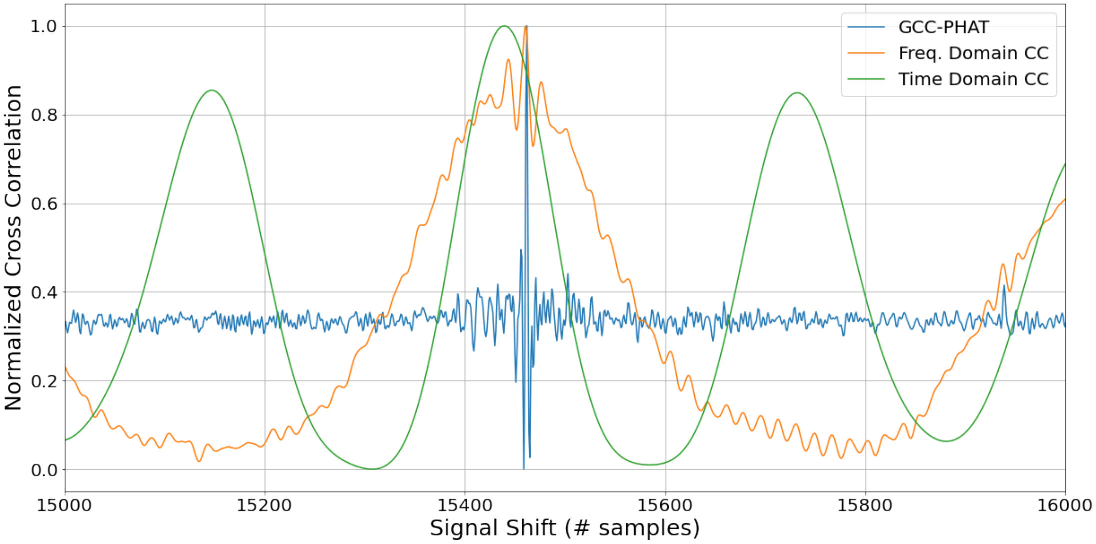

Figure 2.4: Cross Correlation With Different Estimators

that the estimated timing difference is in fact due to the direct speech, and not a stray reflection or reverberation.

The performance of these methods will depend heavily on the chosen audio frame size, background noise and the environment of the robot. Optimization performed in later sections will indicate which method works best for the given tasks.

2.1.2 Generating the Direction Of Arrival

Once the timing difference has been determined, a geometric model is used to estimate the direction of arrival of the sound source. A simplified description is presented in Fig 2.5, showing two microphones M1 and M2, separated by a distance D, with two unique path lengths X1 and X2 to a sound source S.

Given the right triangle made by M1, M2 and the path length X1, with an angle of θ, the opposite then becomes $Dsin\theta$, given the distance D between the two microphones. This distance represents the extra distance the wavefront must travel to reach M1 once it has reached M2. This distance is directly computed from the timing difference τ, and so the measured quantities are related as in Eq (2.7).

$$Dsin\theta = \tau v_{sound} \tag{2.7}$$

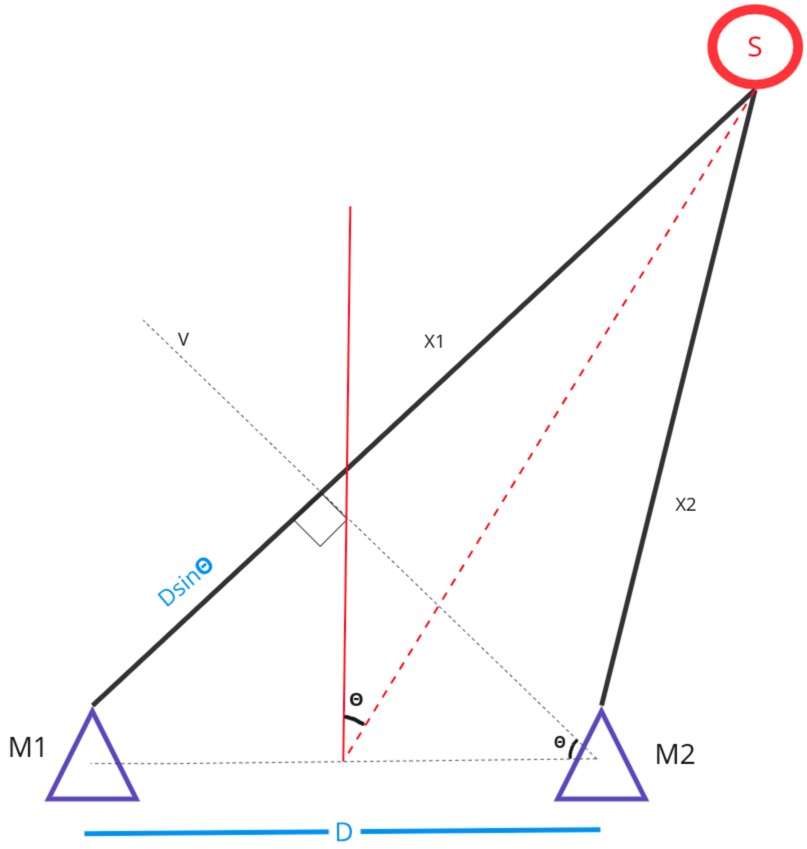

Figure 2.5: Simple DOA geometry

Since the geometry of the robot head setup is not exactly as in this simplified model, the timing difference is used in the Woodworth-Schlosberg model [33] to estimate the DOA on a spherical robotic head, such as the REEM-C's. Eq (2.8) shows the modification that now maps the timing differences τ as a function of the DOA. This new model accounts for the extra radial distance the wavefront must travel to reach the microphone on the other side of the head. This mapping is used to find the corresponding value of the direction of arrival θ in real-time. The ear-to-ear distance D of the REEM-C is calibrated by measuring the ITDs at a number of known angles, and calculating the distance that would result in these measurements. With this method, he average ear-to-ear distance is computed as d = 0.255m.

$$\tau(\theta) = \frac{D}{2 * v_{sound}}(\theta + sin(\theta)) \tag{2.8}$$

2.1.3 Environmental considerations

In a real-world application, it is likely that a humanoid robot will interact with humans in an environment that has some background noise. Voice activity detection (VAD) is a common problem in audio processing contexts, where the goal is to identify when speech is present in an audio recording. When computing the DOA on the REEM-C, the streamed audio will contain a variety of sounds that may not be speech, such as robot operation noises and ambient noise. Reverberant echoes from a talker will also generate spurious DOA estimates as the direction of these echoes is not the same as that of the direct path from the source. We explore two methods to classify if a DOA estimate is ”good” (i.e., the estimate is likely correspond with the direct path of speech from a talker).

Power Onsets

If the microphone signal is split into frames with some window length, the energy in each frame will vary over time based on the fluctuations from the sound source. In a reverberant environment, when a talker stops speaking, it will take some time for the sound energy to decay. When a talker begins speaking, the sound from the direct path will arrive at the microphone before later reflections. Thus, a frame that was more energy than the previous frame (i.e., an onset) is more likely to have relatively more energy from the direct path than a frame that have less energy than the previous one. Here we choose to use successive frame power ratios rather than differences, where an onset is detected if the power ratio exceeds a certain threshold. For a certain frame F_i,

$$F_i = \begin{cases} \text{speech frame} & if\ \delta_{high} > \frac{1}{N}\sum_j^N \frac{F_{i,j}^2}{F_{i-1,j}^2} > \delta_{low} \\ \text{non-speech frame} & else \end{cases} \tag{2.9}$$

The parameters δ_{low} and δ_{high} can be tuned and will depend on the environment of the recording. δ_{low} indicates a minimum required change in frame power, and δ_{high} establishes an upper limit to discriminate against very loud sounds, such as a crashing chair or slammed door. Hence, direct human speech is considered to be limited within a range of power onset values.

Speech-Reverberant Modulation Ratio

The speech-reverberant modulation ratio (SRMR) [38] is a metric that was developed towards predicting the intelligibility of speech in a given audio frame. Conceptually, anechoic speech (i.e., the direct signal) should have significant amplitude modulations between 4-16 Hz, which are related to the acoustic signals that correspond with syllables and phonemes. In the presence of reverberation, delayed and attenuated versions of this acosutic signal are summed together. This results in an increased level of envelope fluctations at higher frequencies. Thus, a ratio of the modulations at low frequencies vs. those at higher frequencies provides a measure that is related to energy of the direct signal vs. that of the reverberant components.

We apply our own lightweight implementation of the SRMR, by first using the Hilbert transform to extract the envelope of the speech frame. The frequency content of the envelope is analyzed by computing the ratio of energy present in modulation bands associate with speech and modulation bands associated with reverberant audio content. The frequencies and bandwidths for the speech and reverberant bands are specified in [38]. Overall the frame classification is performed as follows for a frame F_i,

$$F_i = \begin{cases} \text{speech frame} & if\ \delta_{high} > \frac{\sum_{j=1}^{4} e_j}{\sum_{j=4}^{8} e_j} > \delta_{low} \\ \text{non-speech frame} & else \end{cases} \tag{2.10}$$

where e_j is the energy present in the j-th frequency band of the extracted envelope. This ratio is used as a potential measure for voice activity, and is given thresholds δ_{low} and δ_{high} for similar reasons as the power onsets.

2.2 Problem Statement

A number of methods have been introduced to perform the signal processing necessary for DOA estimation. These methods also involve numerical parameters, which will need to be selected for the human-robot interaction (HRI) task at hand. There is a need to identify the best parameters specifically for a binaural DOA setup on the REEM-C Humanoid Robot, which may be used in reverberant environments for the purposes of HRI. There is also a need to evaluate the implications of using these parameters in real-time, in terms of their accuracy and latency when it comes to HRI scenarios.

This work aims to tackle this problem by presenting a method to identify the best parameters for DOA estimation, including the classification methods, and numerical parameters such as frame sizes and thresholds. Parameters are optimized using a brute force and a Bayesian optimization approach, and used in a real-time implementation on the REEM-C, with a consideration for latency and potential applications for HRI.

2.3 Data Preparation

The binaural DOA setup is deployed onto the robot with a taut headband that places the microphones the head of the REEM-C at the positions that would correspond with human ears, providing a realistic and human-like appearance and configuration. A Scarlett 2i2 audio interface was used with 2 lavalier microphones. This set up was chosen as it is inexpensive and could be adapted and deployed to a wide range of robotics platforms.

Audio recordings were made in a lab environment with the robot operational. This simulates the noise that would be encountered while human-robot interaction scenarios are underway. The annotated periods of speech as well as ground truth locations of the speakers were used to properly estimate the parameters of the DOA pipeline in later sections. The dataset involves 8 recordings to fit parameters and 3 recordings to test the results. Recordings were made in a variety of conditions: stationary vs. moving human talker, while the robot was stationary or performing certain pre-defined motions such as gestures with the arm, head or torso. Other non-speech sounds may also be present, such as foot steps, shifting of chairs and tapping of lab tools against table surfaces. As explained further in the optimization approaches, a weighting is applied to the training set to favour better performance on recordings with more difficult acoustic conditions. The specifics of each recording are shown in Table 2.1.

A good set of parameters will result in the classification that ignores the non-speech sounds but still accurately estimates the DOA of the human talker when they are speaking, even during the relatively noisy operations of the robot.

Fig 2.6 shows the spectrogram of recording 5, with simultaneous robot and speech sounds. The REEM-C performs some motions with the arms, which clearly show up in the spectrogram.

2.4 Optimization Approaches

The optimization and parameter selection takes place via two methods, each with benefits and drawbacks. The results from both methods are compared and contrasted to

Table 2.1: Key Information Of Collected Recordings

RECORDING	**DATASET**	**KEY NOTES**
1	TRAIN	stationary, only speech sounds
2	TRAIN	mobile, only speech sounds
3	TRAIN	stationary, speech + non-speech sounds
4	TRAIN	stationary, simultaneous speech + non-speech sounds
5	TRAIN	stationary, simultaneous robot + speech sounds
6	TRAIN	mobile, speech + non-speech sounds
7	TRAIN	mobile, only speech sounds
8	TRAIN	stationary, speech + non-speech sounds
9	TEST	mobile, robot + speech sounds
10	TEST	stationary, robot + speech sounds
11	TEST	mobile, only speech sounds

assist with choosing the best set of parameters to use on the test set and the final robot implementation.

2.4.1 Brute Force Grid Search

The brute force method attempts every possible combination of parameters across the entire search space and chooses the parameters that best minimize the objective function. From Table 2.2, the brute force method covers 2 classification methods, 3 timing difference methods and a series of numerical values (audio frame sizes, thresholds). To reduce computational time, trials are ended when no windows of speech are found, leading to empty DOA predictions. This is a computationally expensive approach since every possible combination of parameters will need to be attempted, and results will depend on the granularity of the defined search space.

2.4.2 Tree Structured Parzen Estimator

The Tree Structured Parzen Estimator (TPE) is a Bayesian optimization approach that evaluates past results to generate a probabilistic model of the hyperparameters and associated score. Given a series of objective values, *score*, with their respective parameters,

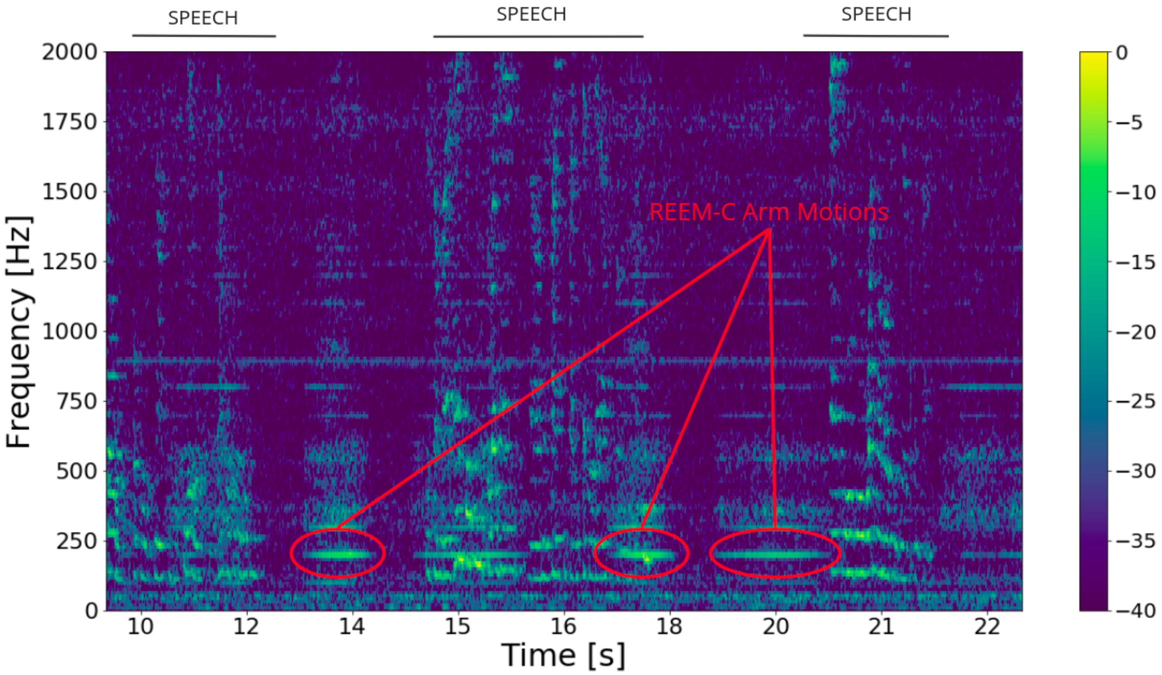

Figure 2.6: Spectrogram of audio sample with human speech and REEM-C motions

$parameters$, the TPE method generates two probability distributions by segmenting the results based on a threshold $score^*$.

$$p(parameters|score) = \begin{cases} l(parameters) & if\ score < score^* \\ g(parameters) & if\ score \geq score^* \end{cases} \tag{2.11}$$

The method then selects parameters with a greater probability of being under $l(parameters)$ than $g(parameters)$, given that $l(parameters)$ is built from trials with more favourable objective values. This informed reasoning is used to select the next set of hyperparameters while updating the two distributions, allowing the method to find an optimal set of parameters while not exhaustively searching the entire parameter space. The TPE is implemented via the hyperopt package [5].

Given the parameter space in Table 2.2, the key difference from the brute force method is that the numerical parameters (frame size, step size, low and high thresholds) are now placed on a continuous distribution. The uniform distribution for the frame size and step size ensure each value has an even chance of being selected. The normal distribution parameters for the low and high thresholds are chosen based on a few tests that may indicate where good thresholds may lie, given that the high threshold must be larger than the low threshold. Table 2.2 outlines the parameter spaces searched for both methods.

2.4.3 Objective Functions

In order to properly define this optimization task, key variables are first defined.

$$\begin{aligned}
&\sigma = [1, 1, 2, 2, 3, 2, 1, 3] \\
&f1(\omega) = 2 * \frac{precision(\omega) \cdot recall(\omega)}{precision(\omega) + recall(\omega)} \\
&mse(\omega) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i(\omega))^2
\end{aligned} \tag{2.12}$$

The vector ω represents all input parameters for any single trial. which come from Table 2.2. The f1-score and mean squared error are then computed for a single trial with a set of parameters ω as per Eq 2.12. The vector σ is a set of weights to compute the weighted average of the metrics generated for the data. We aim to weigh the more complex scenarios higher than the simpler scenarios, and so trials for when extra non-speech sounds are included have a weight of 2, and trials with robot motions occurring throughout the recording have a weight of 3. We believe this weighted average will generate parameters that are better tuned to more complex auditory scenes, as opposed to a standard weighted average across the trials, where good performance in simpler scenarios may dominate the reported metrics.

The objective function used in the optimization approaches varies for each problem. The performance of the DOA estimate classification must be good, and as a result of

Table 2.2: Parameter Spaces Defined For Both Methods. $\mathcal{U}$(min, max)= uniform distribution. $\mathcal{N}$(mean, std) = normal distribution.

Parameter	**Brute Force**	**TPE**
Voice Method	SRMR, PO	SRMR, PO
Timing Method	cross-corr, gcc-phat, gcc-scot	cross-corr, gcc-phat, gcc-scot
Frame Size	(0.1, 1), step = 0.05	$\mathcal{U}$(0.1, 1)
Step Size (%)	(0.1, 1), step= 0.05	$\mathcal{U}$(0.1,1)
Low Threshold	(1,10), step = 0.1	$\mathcal{N}$(3,3)
High Threshold	(3,14), step = 0.1	$\mathcal{N}$(10,3)

potential imbalances in the dataset, the f1-score for classification is considered the metric to optimize. Hence the objective function is formulated for classification as follows, which computes the f1-score for every j-th trial, and aims to minimize the negative of its weighted average.

$$\begin{aligned} \min_{\omega} \quad & -\frac{\sum_{j=1}^{8}(\sigma_j * f1(\omega))}{\sum_{j=1}^{8}(\sigma_j)} \\ \text{s.t.} \quad & 0 < \delta_{low} < \delta_{high} \end{aligned} \tag{2.13}$$

For DOA estimation, the mean squared error is considered as the objective to minimize as the generated estimates and ground truth are continuous variables. The objective for DOA estimation computes the weighted average mean squared error across every j-th trial.

$$\begin{aligned} \min_{\omega} \quad & \frac{\sum_{j=1}^{8}(\sigma_j * mse(\omega))}{\sum_{j=1}^{8}(\sigma_j)} \\ \text{s.t.} \quad & 0 < \delta_{low} < \delta_{high} \end{aligned} \tag{2.14}$$

We also explore how to perform both optimizations at once in a joint manner. The joint optimization aims to minimize the MSE for DOA estimation and maximize F1 for classification. The objective for this method is formulated accordingly, using the two metrics as a fraction. Eq (2.15) shows this formulation.

$$\begin{aligned} \min_{\omega} \quad & \frac{\sum_{j=1}^{8} \sigma_j * \frac{mse(\omega)}{f1(\omega)}}{\sum_{j=1}^{8}(\sigma_j)} \\ \text{s.t.} \quad & 0 < \delta_{low} < \delta_{high} \end{aligned} \tag{2.15}$$

A modification is added to regularize the frame size γ during the optimization. Theoretically, this should result in lower frame sizes found with good results on both DOA and classification tasks, meaning potentially lower latencies when used on the robot for real-time operation. The value of λ is set to 0.5 for this work. This objective function will be helpful to investigate the effect of frame sizes on the final results. Eq (2.16) shows this regularized formulation.

$$\begin{aligned} \min_{\omega} \quad & \frac{\sum_{j=1}^{8} \sigma_j * \frac{mse(\omega)}{f1(\omega)}}{\sum_{j=1}^{8}(\sigma_j)} + \lambda|\gamma| \\ \text{s.t.} \quad & 0 < \delta_{low} < \delta_{high} \end{aligned} \tag{2.16}$$

2.5 Results

We present results for classification, DOA and the joint performance with both parameter search methods. Qualitative evaluation was also conducted on the chosen parameters, and other considerations not included in this optimization are discussed.

Initial quantitative results are presented via contours for visualization. Since there are a total of 6 dimensions to this problem, not all trends can be visualized. These results are further explored below in table form as well.

2.5.1 Brute Force Method

DOA Accuracy

The brute force method results are shown in Fig 2.7 comparing the frame size and step size to the weighted average MSE as a contour plot. The minima, shown as dark regions, occur primarily with larger frame sizes and larger step sizes. The performance on the DOA tends to worsen as the frame size or step size are reduced, indicating that the best choice for this task may require larger audio chunks when used in real-time.

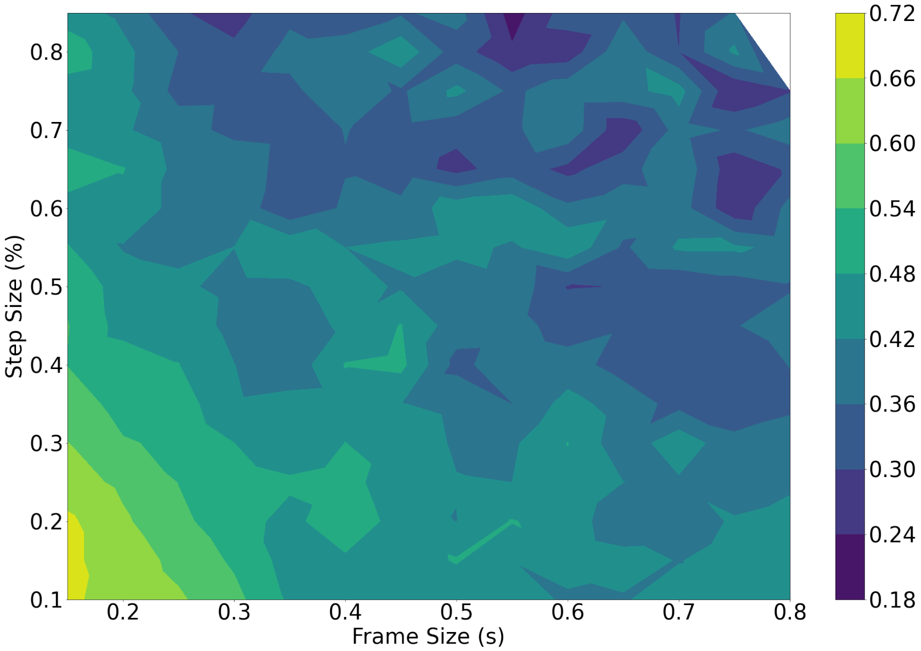

Figure 2.7: Brute force DOA performance against frame size (s) and step size (%)

A tendency towards a higher step size also indicates that it is less important to capture overlapping audio information, as the subsequent ITD estimation is still able to generate good results.

Classification Accuracy

Classification results with power onsets do not exceed an f1-score of 20%, whereas the SRMR performs far better, giving maximum results 70%. The relationship between the thresholds and the classification performance is simple to interpret, as the best results are consistently obtained with a low threshold around 1.5. The high threshold appears to be less important, and can be set to 7 to get good classification results. Fig 2.8 demonstrates the relationship of the classification performance to the set thresholds. The best results are generated using the SRMR as the classification method, with a clear maxima around the specified low threshold of around 1.5 and a high of around 7. Parameters that generated no results, as they detected no windows of speech, show up as white regions in the contours.

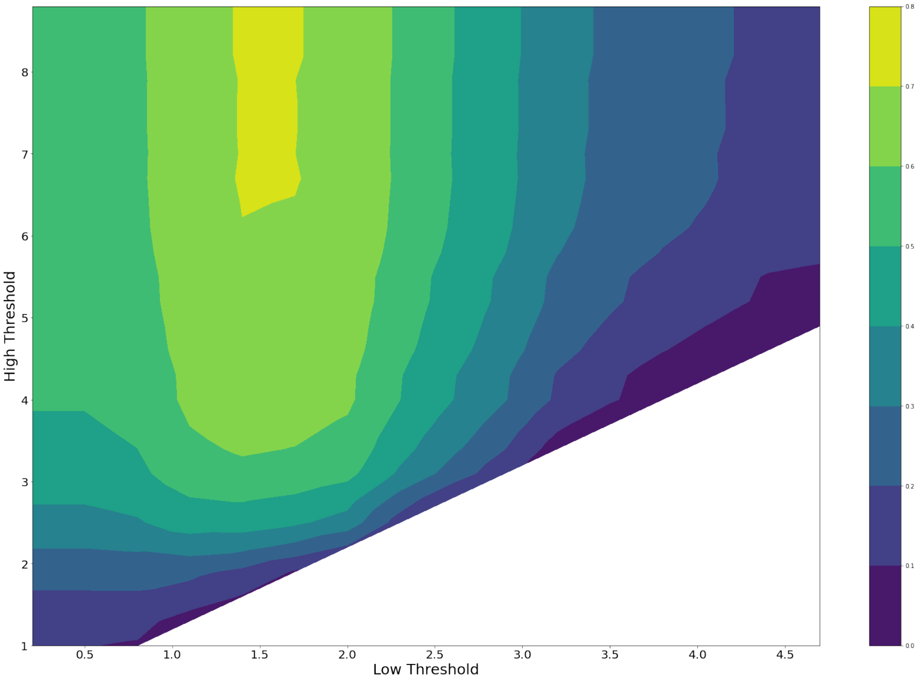

Figure 2.8: Brute Force classification performance against low and high thresholds

2.5.2 TPE Performance

The TPE method is evaluated on all objective tasks next. The TPE method is run for 1000 iterations and completes within a few minutes for each case, highlighting the computational efficiency of this technique while still searching the parameter space in an informed manner.

DOA Accuracy

The results on the DOA task are shown in Fig 2.9 as a contour plot.

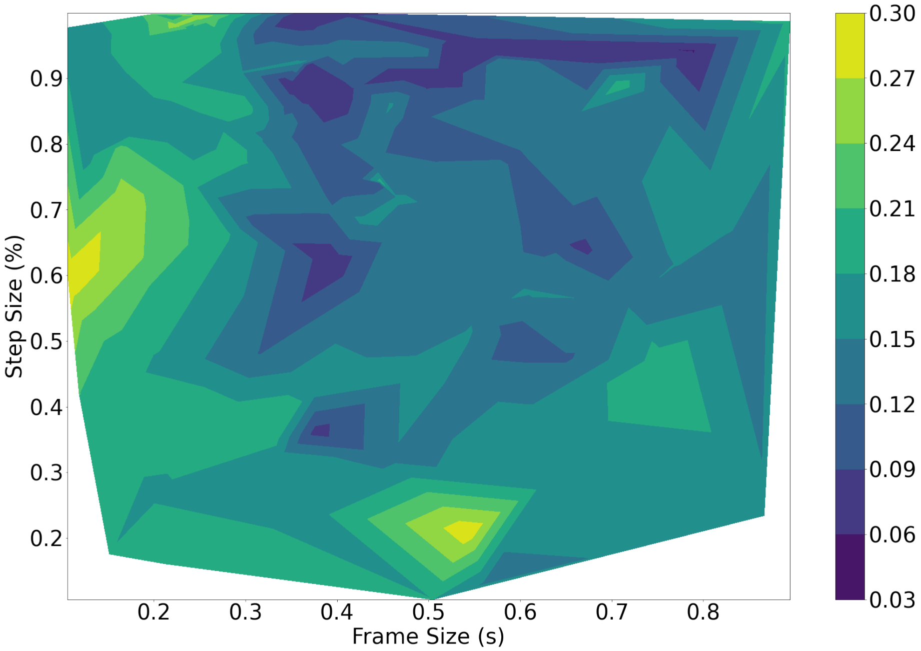

Figure 2.9: TPE DOA performance against frame size (s) and step size (%)

Naturally, TPE performs far fewer iterations and thus outputs fewer data points to visualize. The best results are consistently found using the GCC-PHAT as a timing difference method, and a mixture of the SRMR and power onsets as classification methods. The contour plot shows lower MSE values for larger frame sizes and step sizes, similar to the brute force results. Certain frame sizes and step sizes are never sampled by the estimator since they do not indicate a high probability of generating a good score, leaving white areas on the contour.

Classification Accuracy

We run the Bayesian optimizer for 1000 trials to optimize the task of detecting the presence of speech. The results are shown against the frame size and step size.

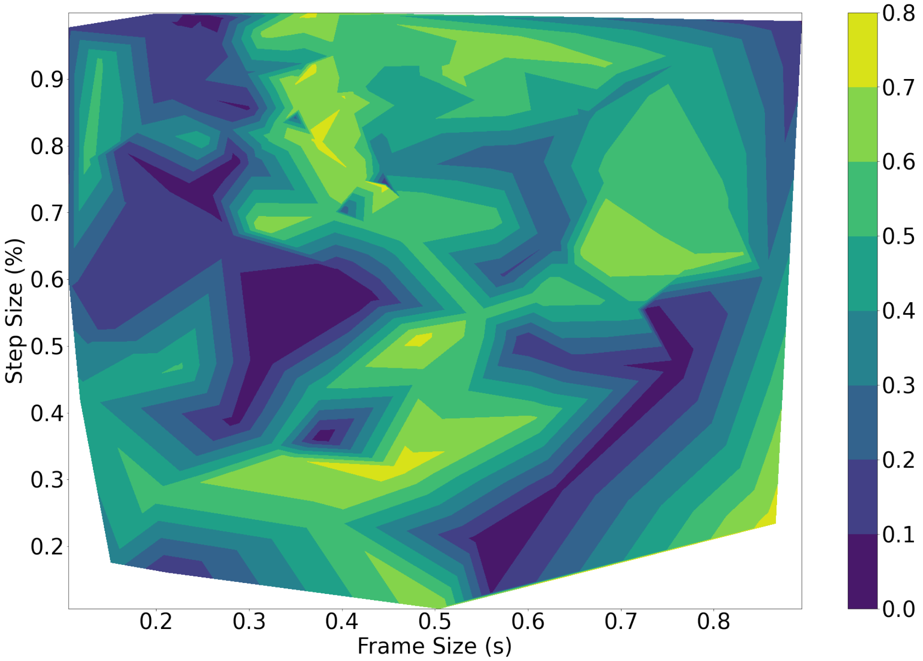

Figure 2.10: TPE classification performance visualized against frame size (s) and step size (%)

A number of maxima in the contour plot are seen with an f1-score around 70%, which are generated using the SRMR as the main method for classification. The optimizer is unable to find good classification results for the power onsets, as the maximum f1-score is around 20%. This is unsurprising, as this metric will only select a good frame if its power exceeds the previous frame's power by a certain factor. For periods of continued speech, the subsequent frame-to-frame power ratio will not be very high, and so a large amount of audio frames containing speech will be rejected.

The relationship of the frame size and step size to the classification performance is more difficult to establish, as compared to the thresholds in Fig 2.8.

Joint Optimization Results

The individual tasks generate different results for the best set of parameters ω_{best}. In order to accomplish both tasks effectively, the joint objective function will need to be optimized. The joint objective function is run through the same TPE optimization pipeline for 5000 iterations.

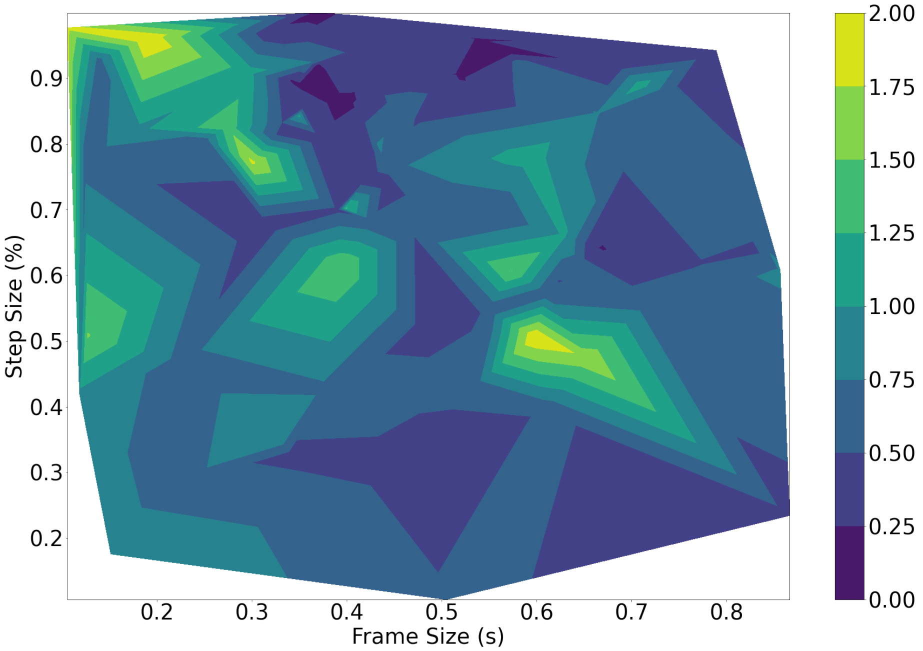

Figure 2.11: Joint Objective Loss vs. Step Size and Frame Size

The best loss values are generated for frame sizes around 400 and 550 ms while optimizing for both the DOA and classification performance. Results are consistently best using the SRMR and GCC-PHAT. The GCC-SCOT appears sparsely in the full results, indicating that the optimizer does not find this technique to be as effective as the GCC-PHAT, and so does not tend to apply it during the learning process. These results tend to agree with what is seen in Fig 2.7 and Fig 2.9, as the minima occur with large step sizes. However, the joint optimization prefers some smaller frame sizes, indicating that optimizing for the classification as well changes the results of the pipeline.

In the context of real-time performance on a robot, larger frame sizes will require more time to generate a response for reorientation by the robot. In order to provide a realistic human-robot interaction, the system should be able to detect and respond within 200-300 ms. Therefore, it is desirable to achieve good classification and DOA performance with lower frame sizes. The optimization is performed again via the TPE method with the regularized objective function, and generates the results as in Fig 2.12.

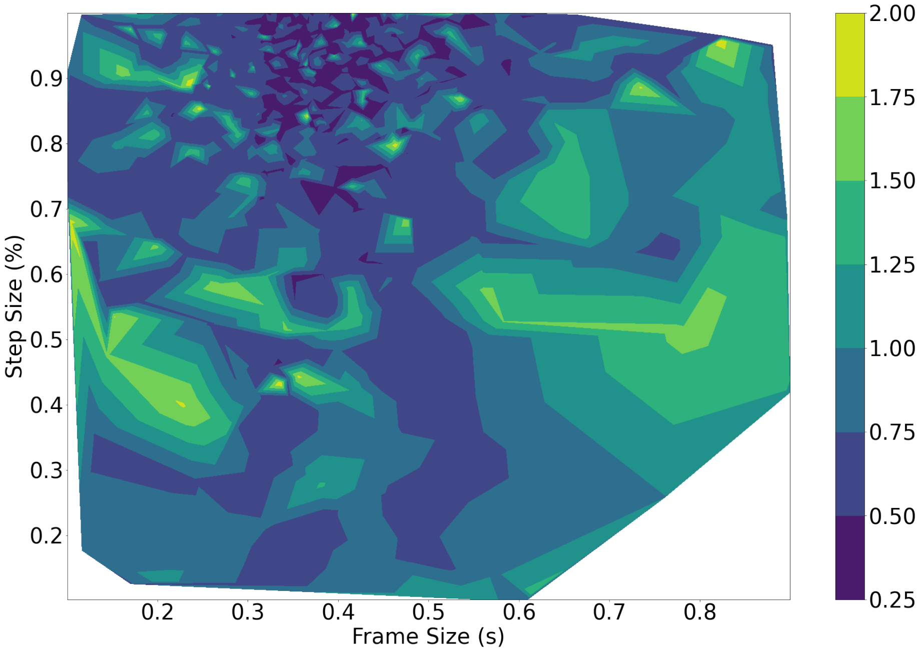

Figure 2.12: Joint Regularized Objective Loss vs. Step Size and Frame Size

As per the contours, more minima are concentrated around the 300-400 ms range for frame sizes. The previously chosen frame sizes above 400 ms are now no longer producing minimal objective values.

Further results in Fig 2.13 show how the average objective values change with regards to the frame size, for the regularized joint objective. With no regularization in the learning process, the larger frame sizes at 500, 700 or 800 ms tend to have lower objective values, which corresponds to previous results. With regularization added, lower frame sizes are favoured, leading to a lowest average objective value at a size of 350 ms.

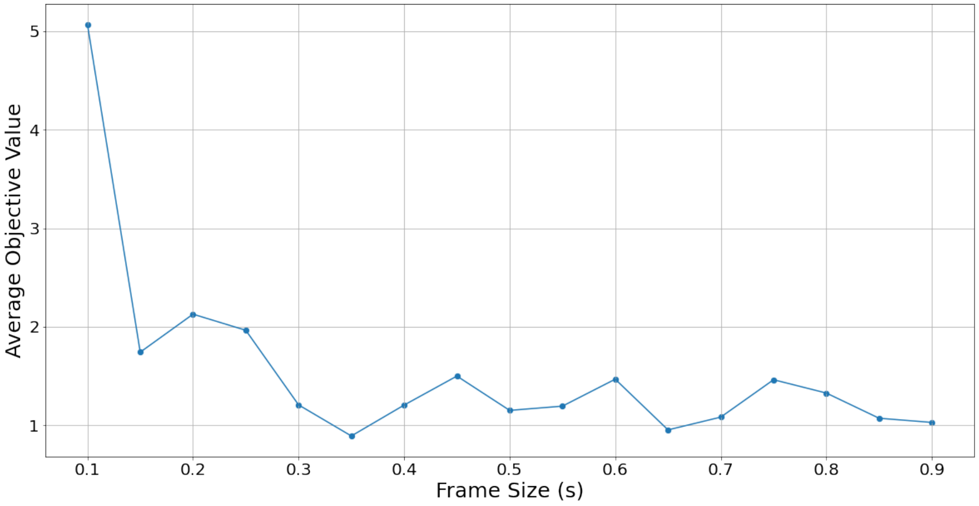

Figure 2.13: Frame Size vs. Average Joint Regularized Objective Loss

The results for all the optimization tasks with either method are shown in Table 2.3. The frame size and step size are reported as they are more crucial to the operation of the robot in real-time. The previous classification results indicate that a good selection for the low and high thresholds is 1.5 and 7, respectively. Overall, the choice of these thresholds is less consequential as their value will not affect the latency of the robot when used in real-time.

Table 2.3 also shows the MSE and F1-score results when different metrics are minimized. For instance, when looking for the best joint objective value, the TPE method yields an MSE of 0.07 and an F1-score of 0.66. In contrast, when looking for the best classification performance, the F1-score is 0.77, with a much higher MSE of 0.15.

It is important to note that the brute force method is limited in its search as it can only evaluate discrete numerical parameters, whereas the TPE method can choose values from continuous distributions. This will have an effect on how the TPE method learns. For the sake of interpretation and evaluation, the best frame size of 339 ms was adjusted to 340 ms, and the step size was adjusted to 0.90 rather than 0.92.

Finally, these best parameters from each method are applied to the test set and generate results as in Table 2.4.

An example of the test set results are shown in Fig 2.14, 2.15 and 2.16 for the 3 recordings.

Table 2.3: Best Parameters And Results For Each Task Across Optimization Methods

Task	DOA	Classification	JOINT	JOINT + REG
Brute Force				
Voice Method	PO	SRMR	SRMR	SRMR
Timing Method	GCC-PHAT	N/A	GCC-PHAT	GCC-PHAT
Frame Size	600	250	550	400
Step Size	0.85	0.10	0.75	0.15
MSE	0.04	0.69	**0.09**	**0.09**
F1-Score	0.11	0.68	**0.61**	**0.68**
TPE Optimizer				
Voice Method	PO	SRMR	SRMR	SRMR
Timing Method	GCC-PHAT	N/A	GCC-PHAT	GCC-PHAT
Frame Size	790	380	572	339
Step Size	0.94	0.35	0.93	0.92
MSE	0.04	0.15	**0.07**	**0.06**
F1-Score	0.14	0.77	**0.66**	**0.72**

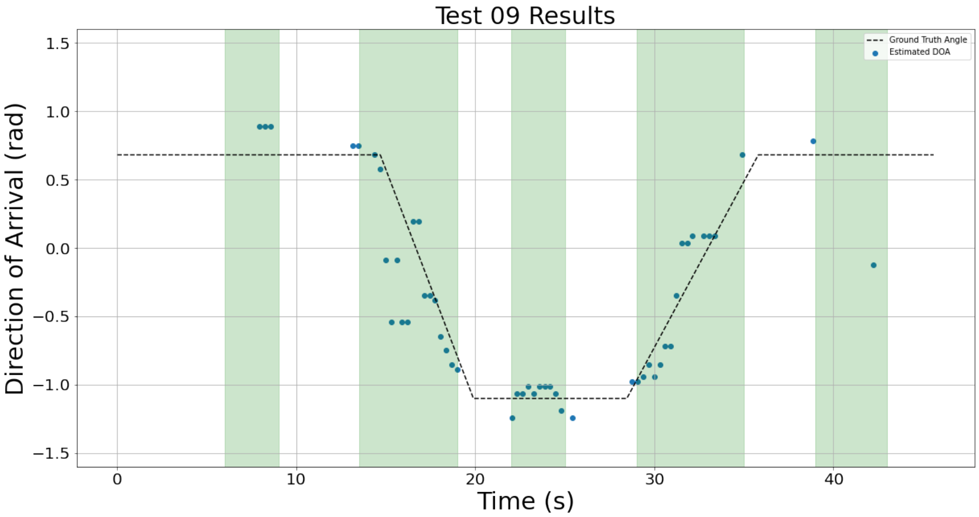

Figure 2.14: Test 09 DOA results. Green = annotated periods of speech. Blue dots = measured DOA. using ω_{best}. Dotted line = ground truth.

Table 2.4: Test Set Performance

Test	MSE Brute Force	MSE TPE	F1-Score Brute Force	F1-Score TPE
Test 9	0.231	0.122	0.844	0.835
Test 10	0.040	0.009	0.855	0.754
Test 11	0.139	0.036	0.841	0.825

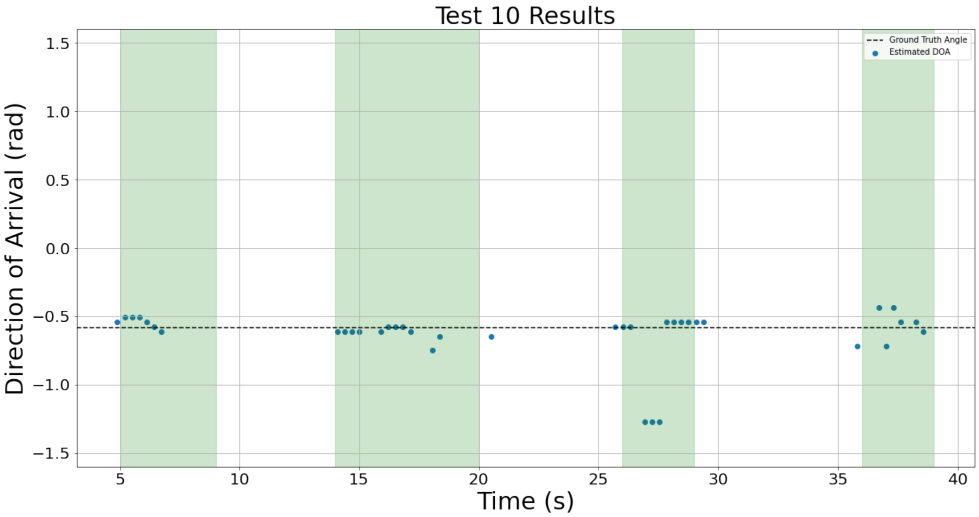

Figure 2.15: Test 10 DOA results. Green = annotated periods of speech. Blue dots = measured DOA using ω_{best}. Dotted line = ground truth.

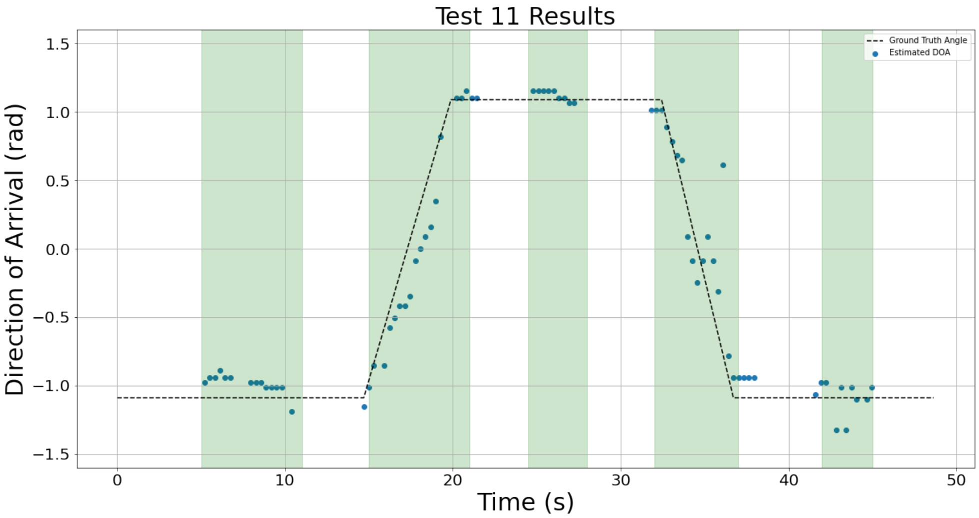

Figure 2.16: Test 11 DOA results. Green = annotated periods of speech. Blue dots = measured DOA using ω_{best}. Dotted line = ground truth.

Given the presented results, and the quantitative results shown in Table 2.3, the best parameters for ω_{best} are found as in Table 2.5.

Table 2.5: Best Overall Parameters ω_{best}

Classification Method	**ITD Method**	**Frame Size**	**Step Size**	**Low Threshold**	**High Threshold**
SRMR	GCC-PHAT	340	0.90	1.5	7

2.6 Use on real robot

We deploy this system onto the REEM-C Humanoid for use in real-time using a full ROS integration. Audio frames are saved to a buffer and processed with the parameters generated from ω_{best}, allowing for real-time estimation of DOA. Fig 2.17 shows the REEM-C's head with microphones installed.

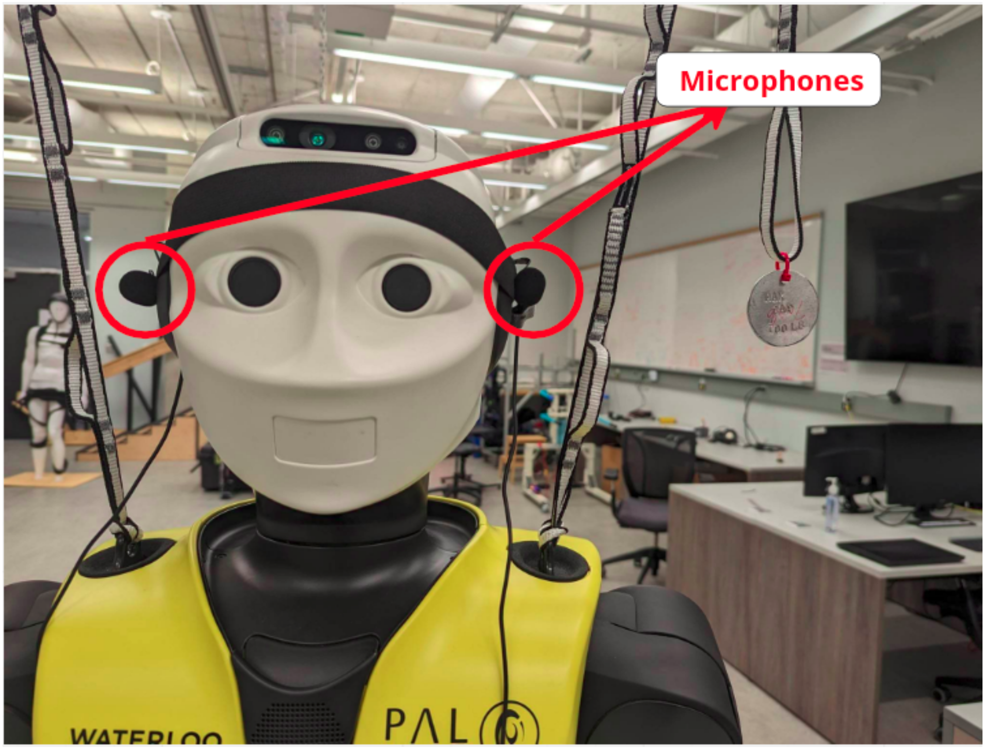

Figure 2.17: Microphone Setup on REEM-C

We also study the effects of latency on the performance of the real-time tracking and estimation. Experiments are carried out with frame sizes of 350ms, 450ms and 600ms, with all other parameters kept constant. The measured DOA are recorded, as well as the audio power level measured by a separate USB microphone, along with the timestamps for both metrics as measured by the ROS network. This allows for identifying how long it takes for a DOA to be measured once the onset of speech has been detected. Fig 2.18 shows the average latency measured with the 3 frame sizes.

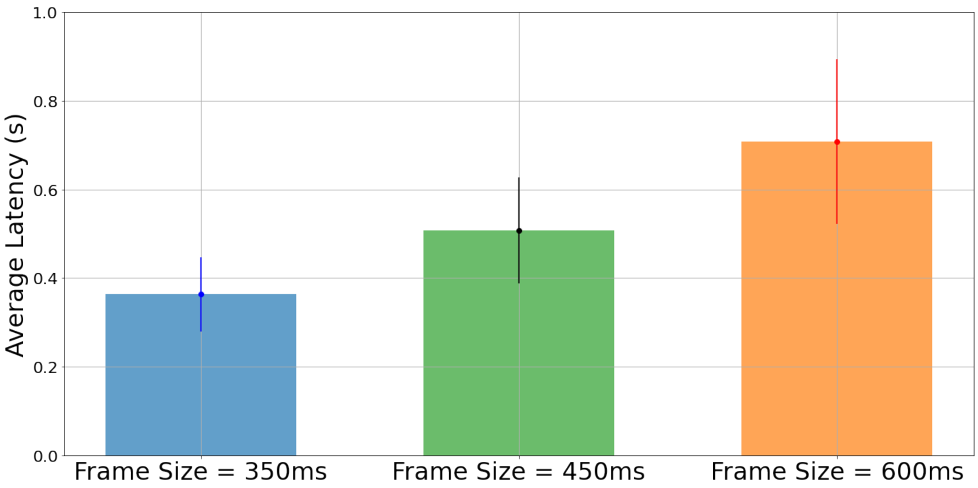

Figure 2.18: Latencies at different frame sizes

The latency is measured to have an average of 0.363s with a standard deviation of 0.076s for the 350ms frame size, an average of 0.508s with a standard deviation of 0.103s for the 450ms frame size, and an average of 0.708s and a standard deviation of 0.169s for the 600ms frame size. For 10 separate DOA measurements at 350ms and 600ms, a two-tailed t-test for their latencies yields a p-value of 0.0019. With α set to 0.05, this indicates that the choice of frame size is indeed significant for real-time use, further validating the regularization applied in the optimization and the choice of lower frame sizes for ω_{best}.

2.7 Discussion

The generated results across all methods have noticeable similarities and differences. For the DOA task, as per Figure 2.7 and Figure 2.9, the brute force and the Bayesian methods mostly lead to frame sizes that are 500 ms or larger, and step sizes larger than 75%. The GCC-PHAT succeeds most often as a method to estimate timing difference compared to the standard beamformer and the GCC-SCOT. The best frame selection method happens to be the power onset. This is unsurprising as power onsets will should be dominated by energy from the direct path (i.e., have a high direct to reverberant energy ratio) and so will likely produce accurate DOA estimates when the power onset condition is met. However, this comes with the trade off of rejecting many other frames containing speech as the subsequent frame-to-frame power ratio during periods of continuous speech can be similar. This potentially ignores many frames where the ratio direct to reverberant energy may still be high.

This is more evident when optimizing for the classification task. Both optimization methods point towards using the SRMR as the main method to detect periods of speech, with classification based on power onsets consistently producing poor results (no more than 0.2 F1-score). In a perfect scenario, the pipeline can identify all present windows containing speech, and would yield an F1-score of 1 for classification. As a result, it would be necessary for the pipeline to correctly detect as many windows of speech as possible. In general, an F1-score above 0.7 would be considered sufficient, as it indicates a strong ability to both detect speech when it is present, and not identify speech when it is not present.

Since both tasks produce different results, the joint objective results should be investigated to identify parameters that perform well for both the classification and DOA tasks. The joint objective task for both methods favours the SRMR and GCC-PHAT for processing the audio frames. In addition, the brute force method resultsin a frame size of 550ms, whereas the TPE method finds the best results to occur with a frame size of 572ms. These results are in close agreement, but require a long latency when implemented on a robot – orienting behaviour on the robot will lag any movement of a talker by half a second.

Studies in turn-taking dynamics and conversational behaviour indicate that humans take on average 200-300 ms to respond to their partners [14], suggesting that lower frame sizes will be more required for more natural for HRI behaviour. The joint regularized task produces the best results with lower frame sizes, as is depicted by where the minima lie on the contour plot in Fig 2.11 and Fig 2.12. The best results for the study are then taken from the joint regularized task using the TPE method. It is important to note that regardless of how the optimization is performed, good results are rarely found for both tasks with frame sizes less than 300 ms, as shown by where the minima lie in Fig 2.12. We suspect that this is due to the calculations involved in computing the SRMR. For the SRMR, the process involves studying the modulation of the speech signal via its envelope, and extracting the energies present in certain bands of this envelope. The lowest frequency band for this metric was centred at 4Hz. Thus, one period of this modulation corresponds with 250 ms. Frame sizes shorter than this length may result in inaccurate estimation of the 4 Hz component of the modulation energy. Thus, the use of SRMR as it is was defined here may impose a minimum latency that is too long to achieve human-like behaviour. While increasing the minimum modulation frequency used in the SRMR would reduce the minimum latency, further work is need to determine the effect this would have on classification performance. If other classification methods are explored, minimum latencies should be less than 200 ms.

Table 2.3 gives further insight to the numerical performance of these methods. The values indicate differences in performance when searching for different objective value minima. Numbers in bold indicate the notable differences in performance when optimizing for either loss value. For instance, using the brute force method, searching for the minimal regularized objective provides the same MSE of 0.09 as found with the unregularized objective, but gives a F1-score improvement to 0.68 from 0.61. Similarly, the TPE method sees both a decrease in MSE from 0.07 to 0.06, and an increase in F1-score from 0.66 to 0.72 when minimizing the regularized objective as opposed to the unregularized objective. This is supporting evidence that the joint objective function was appropriate for this problem as both classification and DOA tasks are performed to an acceptable level with the best parameters ω_{best}.

Furthermore, the regularized objective also provides smaller frame sizes along with the performance gain. This is evidence that a regularized objective was helpful for the TPE in its learning process. However, since the method only evaluates a subset of the parameter space, this may come down to the randomness in its choice of parameters, explaining why the unregularized version did not find similar parameters.

The results on the test set show that these parameters are reasonable and have not overfit on the training set, as the performance on both tasks are good for all 3 test set recordings. We also see that the TPE method generates much better MSE compared to the brute force method, but slightly worse F1-score, as per Table 2.4. We suspect that the lower step size from the brute force method provide a greater resolution for identifying speech on the microphone signals, leading to slightly better classification performance on test data.

Test 9 sees a higher MSE than the other two tests; we hypothesize that this is most likely due to the subject moving farther and closer to the robot as opposed to maintaining a similar distance as in test 10 and 11. This may simulate more realistic human-robot interaction scenarios, and could require some improvements to reflect better results.

With a functioning sound source localization pipeline in real-time, the potential for HRI can be expanded. For instance, if moving conversational partners can be detected by the robot, HRI can be augmented by implementing a human-like, realistic tracking behaviour. Motion capture analysis and modeling of the head, shoulders and feet such as in [2] can be applied for this purpose.

2.8 Conclusions

This work presented a pipeline to perform binaural direction of arrival estimation on a humanoid robot. Optimization procedures were used to improve the performance on a number of trials in the acoustic environment of the robot, and are able to find consistent results regarding the best classification and ITD methods, including the relevant numerical parameters. Test set results indicate that the chosen parameters are appropriate for a variety of acoustic scenarios. A method to use this pipeline on the real REEM-C is also presented, with considerations for real-time latency and performance.

Chapter 3

Vision Systems For Human Robot Interaction

3.1 Introduction

Human-robot interaction systems often involve using either audio, video, or both to facilitate tasks such as working in assembly workspaces [1] or communicating with human subjects [32]. To engage in conversations that are perceived as being natural, there is a need to employ a combination of these systems. A robust set of imaging algorithms are therefore required to allow for this functionality.

Voice activity detection is easily accomplished when audio information of present interlocutors is available. The applications of visual voice activity detection involve scenarios when audio is not available due to privacy or hardware restrictions, or may be unreliable due to a noisy or adverse acoustical environment. Additionally, gaze estimation is important to identify conversational cues and to estimate where interlocutors are directing their attention. Alongside a method to estimate the angle of the interlocutor relative to the robot, a framework to allow for a more complete human-robot interaction can be developed. These problems are often approached with computationally heavy deep learning models trained on extensive datasets [13, 41], and therefore require a lightweight, classical computer vision alternative. [1]

[1] The content of this chapter is from the following conference paper: Barot, P., MacDonald, E., Mombaur, K. (2023). Vision Systems For Identifying Interlocutor Behaviour And Augmenting Human-Robot Interaction. Journal of Computational Vision and Imaging Systems, 8(1), 55–58. https://doi.org/10.15353/jcvis.v8i1.5377

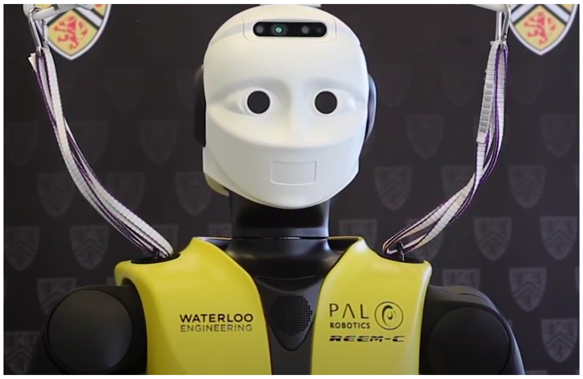

Figure 3.1: REEM-C Humanoid Robot, with RGB-D camera shown on forehead

All systems are designed for the REEM-C Humanoid Robot, which utilizes a RealSense RGB-D camera. Due to bandwidth limitations on the ROS network, these systems are designed for a feed of 15 fps at a resolution of 640 x 480. An image of the REEM-C is shown in Figure 3.1.

Experimental data is collected by recording audio and video simultaneously of human conversations from the REEM-C. Audio is recorded in two channels, at a standard rate of 44.1 kHz, and video frames are recorded directly from the REEM-C camera. The environment in which the data is collected is the research lab where the REEM-C resides, which features variable overhead lighting helping to simulate different scenarios the robot may encounter. No other sounds, other than those from the robot's natural operations, were present during the recordings to maximize the reliability of the recorded data.

3.2 Visual Voice Activity Detection

Conversational scenarios with a humanoid robot will require identifying when humans are speaking and when they are silent. Visually, this can be done by identifying features that correlate to speech. To understand which features are most correlated with speech, audio data is broken into frames of 1/15th of a second, to match the data extracted from video. Features are directly extracted from frames using facial landmarks from the DLIB detector, which is commonly used in the literature [18]. These landmarks allow for measurement of important characteristics from detected faces, such as those of the mouth or eye areas. Figure 3.2 shows detected facial landmarks on a subject, and the extracted mouth and eye

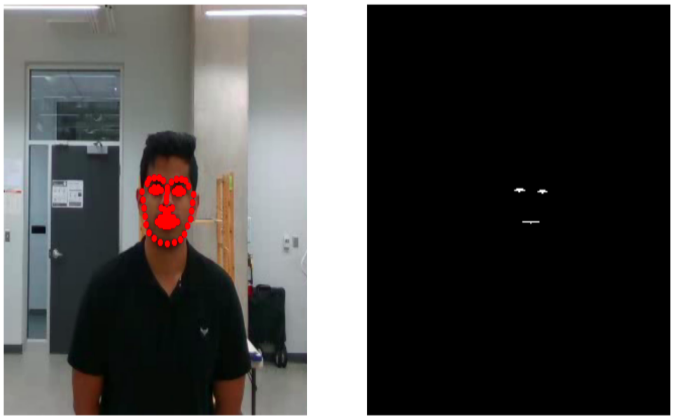

Figure 3.2: Detected facial landmarks shown in red (left), and mouth and eye regions shown as binary masks (right)

regions as binary masks.

Features that may be related to speech are extracted on each frame, which include but are not limited to, mouth height, mouth area, Sobel filter gradients, and HSL information from the mouth areas. A common method to process features includes using a sliding window approach, which moves frame-by-frame, spanning the sequence of collected data. To mimic concepts used in audio signal processing, metrics such as the mean and the power of the window [30] are used to identify behaviour that is of interest. For instance, we suppose that as a person speaks, the area of their mouth will be larger than when they do not speak, which will be reflected in the mouth area signal. The average of the sliding windows is taken for video data, time aligned with audio data, and compared to the audio frame power. For a given window of data, the mean and the power of the window are computed as in Equation 3.1 and Equation 3.2. A window size of N=5 frames is chosen for this experiment.

$$Mean(x) = (1/N)\sum_{n=1}^{N} x[n] \tag{3.1}$$

$$Power(x) = \sum_{n=1}^{N} x^2[n] \tag{3.2}$$

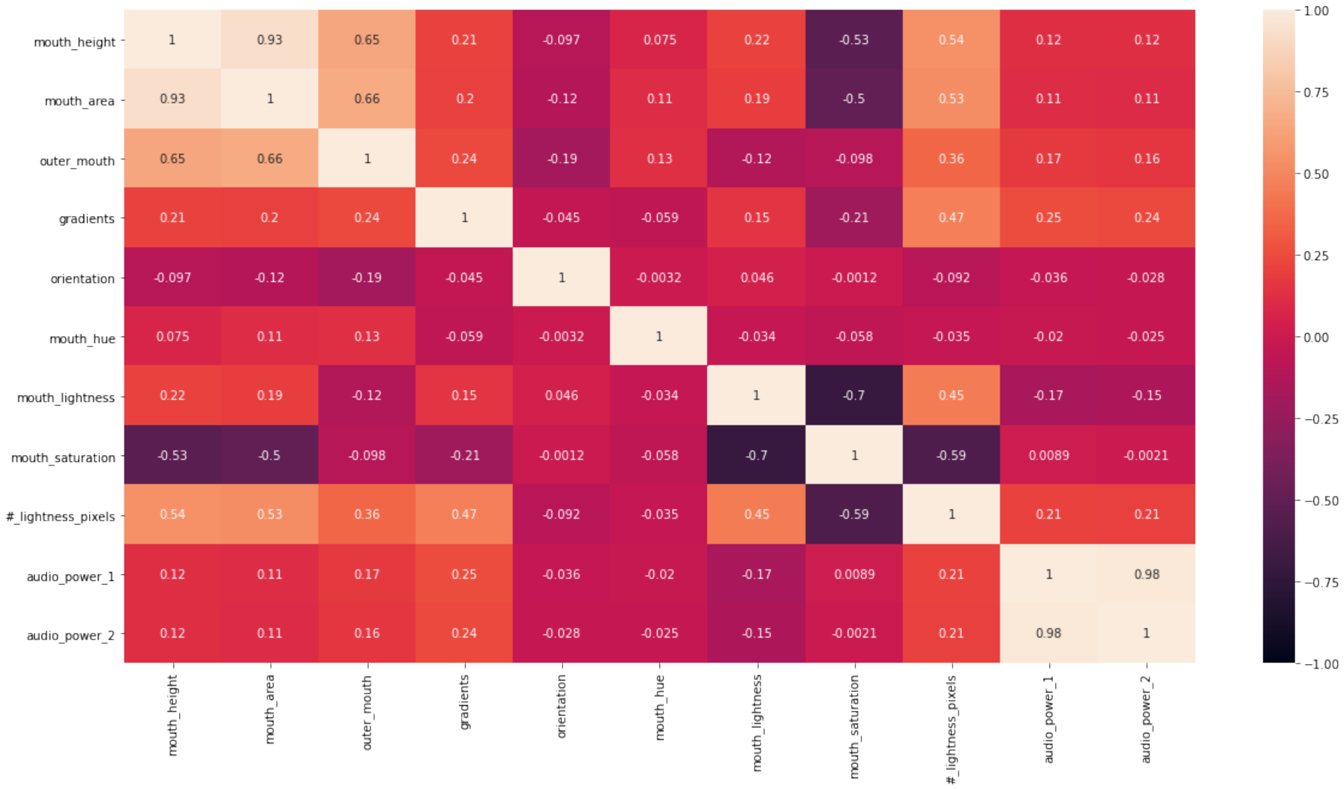

Figure 3.3: Sliding window feature means in correlation matrix with audio frame power

A correlation matrix was generated to better understand which visual features correlate with two channel audio power. As described previously, some of the extracted features include the height and area of the mouth, colouration changes in the mouth and Sobel gradient magnitude and orientation features. These features are normalized by the subject's mouth size, to accommodate for different facial features between subjects. An example correlation matrix is shown in Figure 3.3.

Noticeably, certain features have some correlation with audio power, and may be used in tandem to identify periods of voice activity. Lightness pixels indicates the number of pixels in the mouth region below the average lightness of the mouth area detected in the first frame. This feature's extraction is shown in Equation 3.3, on every frame.

Figure 3.4 and Figure 3.5 show window averages for lightness pixels of the mouth region, and the area of the mouth, respectively. Green windows indicate periods of voice activity.

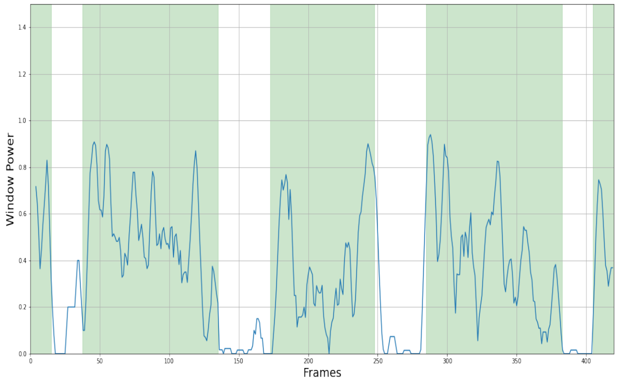

Figure 3.4: Sliding window average of lightness pixels feature. Green = voice activity, White = no voice activity

$$
\begin{aligned}
&\text{lightness pixels} = \frac{\sum_{l=1}^{L} P_l < T}{\text{mouth area}} \\
&P_l = \text{l-th pixel of mouth area, in the lightness channel} \\
&T = (1/L)\ (\text{sum of all pixel lightness in mouth of first frame})
\end{aligned}
\tag{3.3}
$$

The threshold, T, is updated every 10 seconds to account for possible changes in lighting or positioning of the subject. It can be seen that the lightness feature corresponds well with windows of voice activity. A similar pattern is seen when looking at the power of the mouth height measurements. We attribute the fluctuation of values within the periods of voice activity to the variability in the movement of the articulators required to pronounce different phonemes.

To capture the observed change in these features between the intervals with versus without voice activity, a measure is used in Equation 3.4 that is based on modeling the data windows as Gaussian distributions [42]. After computing the mean and power of each frame, voice activity is classified as per the following conditions from [42]. Q is the probability of a Gaussian random variable given the window mean and standard deviation, and PFA (probability of false alarm) is set to 1%.

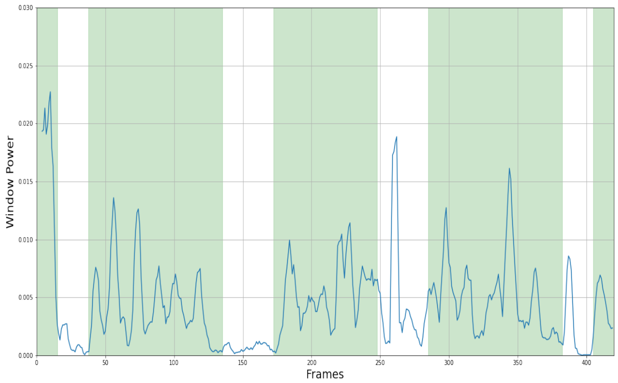

Figure 3.5: Sliding window power of mouth heights feature. Green = voice activity, White = no voice activity

$$
\begin{aligned}
&Mean(x) > \alpha_1 \ and \ Power(x) > \alpha_2 \\
&\alpha_1 = \sqrt{\frac{\sigma^2}{N}} Q^{-1}(PFA) \\
&\alpha_2 = \sigma^2 Q^{-1}(PFA)
\end{aligned}
\tag{3.4}
$$

This algorithm was applied to a test video with two subjects conversing and generated results with the lightness pixels feature as shown in Figure 3.6.

The estimates are dense in periods of speech, indicating the algorithm is able to isolate distinct periods of voice activity for both speakers. However, there are a large number of false positives and some false negatives. The use of other features may be key in reducing the rate of false positives and more accurately identifying the intervals of voice activity when a talker speaks.

This algorithm is also functional online, and accommodates for new subjects entering or leaving the field of view of the robot. Quantitative results are presented for this algorithm in Table 1.

Overall, the quantitative tests of the algorithm produced good results, indicating the algorithm can identify periods of speech among one or more subjects that are simultaneously present in the robot's field of view.

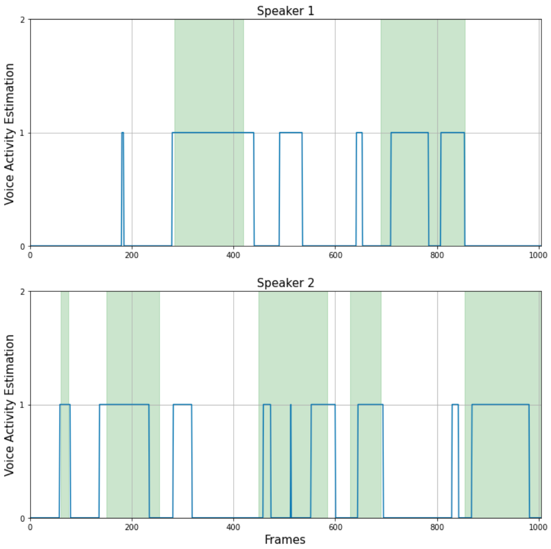

Figure 3.6: Voice activity detection estimates, in blue, visualized against annotated windows of speech for speaker 1 and speaker 2 in conversation. Green = voice activity, White = no voice activity

3.3 Gaze Estimation

Gaze is estimated using binary masks on each half of the eye, for both eyes. For the left eye for instance, a binary mask for all pixels enclosed by points (36,37,41) and (38,39,40) are extracted. The amount of sclera present in either half of the eye indicates where the iris may be placed [47], thereby estimating the direction of the subject's gaze. This amount of whiteness in the eye can be measured via the average lightness in the eye halves in the HSL space. The ratio of lightness in the left half of the eye to the lightness in the whole eye is computed, for both eyes. The same is done for the right half of the eyes, and then the two are subtracted to generate a difference in lightness for both halves of the eyes (see Equations 3.5, 3.6, 3.7, 3.8 and 3.9).

$$\mathrm{L_{eye}, L_{half}} = \frac{\text{avg. lightness in left half of left eye}}{\text{avg. lightness in left eye}} \tag{3.5}$$

$$\mathrm{R_{eye}, L_{half}} = \frac{\text{avg. lightness in left half of right eye}}{\text{avg. lightness in right eye}} \tag{3.6}$$

Table 3.1: Visual Voice Activity Detection Results

Test	Subject	Accuracy	Precision	Recall	F1-Score
1	1	78.5	85.9	81.6	83.7
2	1	86.7	76.1	84.9	80.2
2	2	75.1	75.9	66.9	71.1
3	1	90.7	91.9	91.9	91.9
3	2	84.9	70.9	83.1	76.5

$$\mathrm{L_{eye}, R_{half}} = \frac{\text{avg. lightness in right half of left eye}}{\text{avg. lightness in left eye}} \tag{3.7}$$

$$\mathrm{R_{eye}, R_{half}} = \frac{\text{avg. lightness in right half of right eye}}{\text{avg. lightness in right eye}} \tag{3.8}$$

$$\begin{aligned} \text{gaze ratio} =& (\mathrm{L_{eye}, L_{half}}) + (\mathrm{R_{eye}, L_{half}}) \\ & - (\mathrm{L_{eye}, R_{half}}) - (\mathrm{R_{eye}, R_{half}}) \end{aligned} \tag{3.9}$$

This gaze ratio is extracted on every frame, and checked against determined thresholds to identify if the person is looking to their right, left, or forward. Figure 3.7 shows the gaze ratio for a test video, with annotated windows for where the subject's gaze was directed.

From the test data, hard thresholds are imposed to determine the subject's gaze, as shown in Equation 3.10.

$$\text{gaze direction} = \begin{cases} \text{left} & if \text{ gaze ratio } > 0.5 \\ \text{right} & if \text{ gaze ratio } < -0.5 \\ \text{forward} & else \end{cases} \tag{3.10}$$

3.4 Interlocutor Angle Identification

Allowing for realistic conversations also involves identifying where exactly the speakers are, in terms of angular displacement relative to the robot. This can be done by using the RGB-D camera's depth information, combined with the location of the subject's face in the 2D image, to triangulate their position. The location of the subject's face is taken as

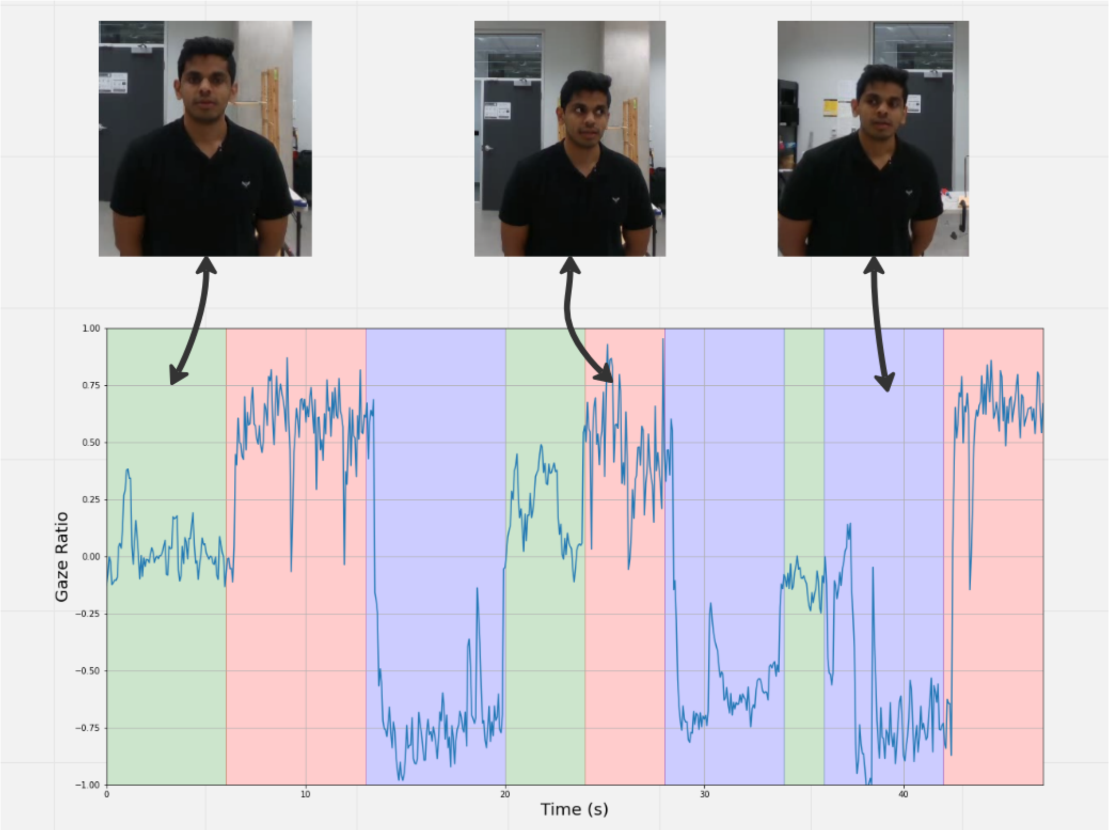

Figure 3.7: Gaze ratio plotted against annotated windows of gaze. Red = gaze to the left, blue = gaze to the right, and green = gaze directed forward

the average of the coordinates of the landmarks that outline the face. This is demonstrated in Figure 3.8.

This vision system opens up the possibility of orienting the robot towards the person who is detected to be talking, in a way that replicates natural human behaviour. For demonstration and analysis, the robot-relative estimated interlocutor angle is converted to to an absolute coordinate system (i.e., room-relative) based on re-orientations of the robot head, torso, and feet.

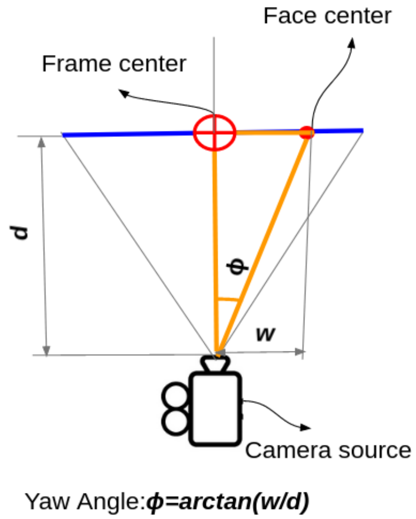

Figure 3.8: Identification of interlocutor angle using depth information

3.5 Discussion

All three visual processing systems, voice-activity detection, interlocutor gaze direction estimation, interlocutor angle identification, must work together to facilitate a more complete human robot interaction scenario. Each human that is present in the scene is characterized on every frame by their voice activity status, their gaze, and their angle relative to the robot. This information is transmitted through the ROS network, which allows for a framework to be developed for interacting with one or more subjects. Future work may explore how these visual systems can be utilized in tandem to create this framework.

This framework provides real-time estimates of cues that are necessary for achieving more natural movements and orienting behaviour during human-robot conversational interactions [36, 25, 7], including gaze following, implementing conversational cues and more. While the methods presented here involve only visual processing, the addition of audio-based processing is likely needed to achieve human-robot interactions that are more robust. For example, algorithms for sound source localization [30] can be used to correct for scenarios where faces are undetected, masked, or outside the field of view for the purposes of voice activity detection. Direction of arrival estimates using audio can also be used to correct for errors in visual voice activity detection, by increasing confidence in estimates of who is speaking, or verifying false positives. Obviously, audio processing is needed to perform speech recognition, which is also required for a robot to hold a conversation.

3.6 Conclusion

In conclusion, this work explored the potential in developing vision based systems for identifying the behaviour of interlocutors with the REEM-C. Three unique imaging systems are proposed to augment human-robot interactions by providing the robotic system greater understanding of its interlocutor's activity. Given the good accuracy on these algorithms, future work may explore their synthesis, to allow for the REEM-C to intelligently interact with multiple humans simultaneously.

Chapter 4

Natural head and body orientation for humanoid robots during conversations with moving human partners through motion capture analysis

4.1 Introduction

By incorporating more human-like behaviour into robots, it should be possible to achieve human-robot interactions that are perceived as more natural. When holding a conversation, people do not always remain stationary. As they move, their conversational partners must also move and reorient themselves to maintain the social norms of holding a conversation. While methods to estimate the the angle between a robot's current orientation and incoming sound sources [30] or people in a visual scene [35], [21] have been developed, humans do not always orient themselves to directly face a conversational partner. In this study, we use motion capture to characterize how humans reoriented themselves during conversation with an interlocutor that moved from time to time. [1]

[1] The content of this chapter is from the following conference paper: Pranav Barot, Ewen MacDonald, Katja Mombaur, Natural Head And Body Orientation For Humanoid Robots During Conversations With Moving Human Partners Through Motion Capture Analysis, In Conference on Advanced Robotics and its Social Impact (ARSO), Berlin, Germany, 2023

Motion capture data helps address such an issue. For instance in a motion retargeting work such as [17], positions and orientations of the head, hands and feet were tracked to retarget the motions of an original character to a computer animation in an online fashion. For works specific to humanoid robotics, [10], [44] measure and model a variety of human motions, and demonstrate them on state-of-the-art humanoid robots. With regards to coordination of body parts, [43] provides analysis of the contributions of head and torso while studying participants' gaze shifts. Works relating to gestures and movement based communication [26], [37] try to improve the human-like quality of robotic motions during human-robot interaction. Dynamic motions of body parts are of significant interest in [15], where head and body orientations are modeled during locomotion. The body and the head are also studied in [16], specifically in scenarios of walking and turning. Human motions are also imitated on humanoid robots using motion capture data [4].

In this work we present a motion capture study for the specific task of tracking moving conversational partners, as an extension to [15]. Human models are reduced to 3 unique segments, the head, the shoulders and the feet, and mathematical models are developed to replicate the manner in which these 3 segments behave during the trials. Validations are performed on developed models and sources of error are discussed as well.

4.1.1 Contributions

The contributions for this study are as follows.

- A motion capture study centered around tracking moving conversational partners
- Mathematical models and parameter estimations using captured data for interlocutor tracking
- Validation of models and humanoid robot implementation

This work aims to add to the existing scientific literature on non-verbal behaviours by also studying the dynamics of a moving conversational partner in the context of human-robot interaction. Section 4.2 explains the design and protocols of the motion capture study. Section 4.3 discusses model design and mathematical approximations. Section 4.4 demonstrates results and section 5.8 discusses their implications. Section 4.6 shows an implementation directly onto the REEM-C Robot. Finally, section 5.9 summarizes the findings of this work and possible improvements and extensions.

4.2 Motion Capture Study Methods

4.2.1 Marker Placements

Given that the head, shoulders and feet are of key interest in this study, markers were placed accordingly to allow calculation of the orientation of each unique segment.

Figure 4.1 demonstrates 3 markers placed on the head of the interlocutor, 2 markers placed on the shoulders, and 2 markers placed on either end of the feet. This allows for estimation of head, shoulder and feet angles of the participant. 2 additional markers were placed on the pelvis but not used in this paper. Markers were also placed on the interlocutor, primarily to confirm their angle relative to the participant.

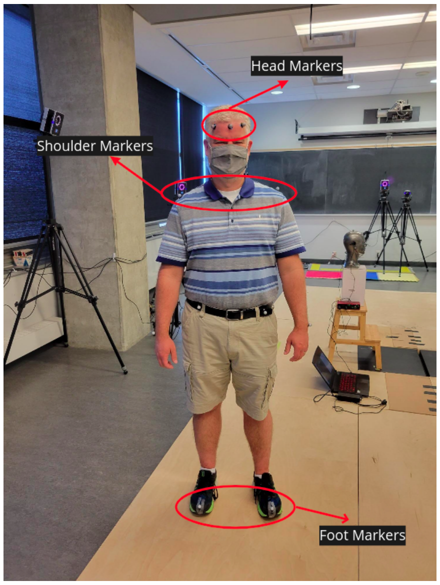

Figure 4.1: Markers placed on interlocutor to measure segment angles

4.2.2 Procedure

The interlocutor led the conversation and engaged in a variety of topics with the participant, such as favourite cuisine, past travel experiences, etc. During this conversation, the interlocutor walked around the motion capture study space. The interlocutor was free to move from $\frac{\pi}{2}$ to $\frac{-\pi}{2}$ radians, as an angle measured from the participant's starting orientation, and maintained a similar distance throughout the conversation (see Figure 4.3 for a representative example). The participant's starting orientation involves them facing directly forward, and was considered to be the reference from which the interlocutor's angle was measured, as seen in Figure 4.6. A positive angle is to the participant's right, and a negative angle is to the participant's left. Participants were instructed to naturally engage in the conversation as they saw fit. They were instructed to remain in the same physical location but could otherwise move as they would naturally.

Figure 4.2 shows an interlocutor engaging in conversation as they naturally reorient themselves to follow the moving participant.

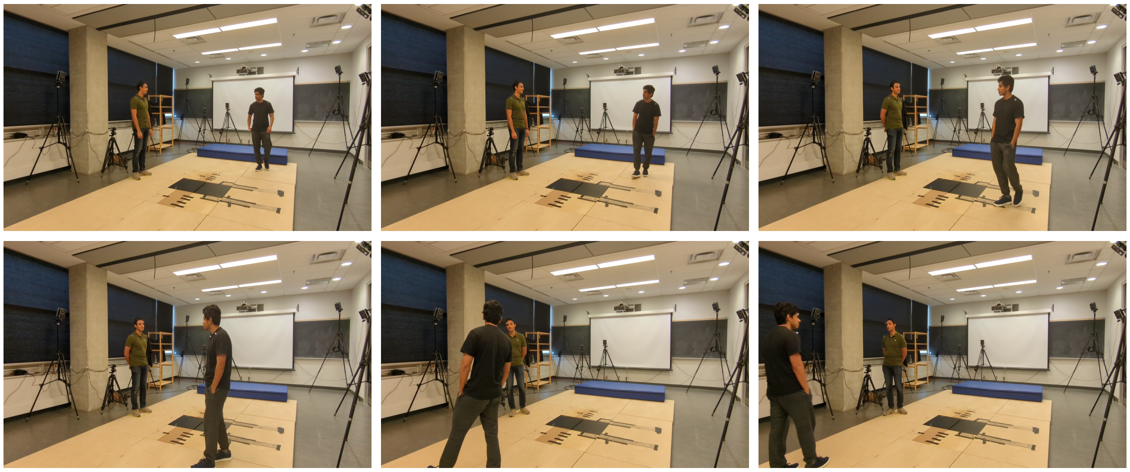

Figure 4.2: Frames depicting the REEM-C moving its head, shoulders and feet to orient to a moving conversational partner (view from top left to bottom right)

A total of 21 conversations were recorded across 5 participants. For this study 5 conversations were discarded due to issues with the motion capture recording, leaving a total of 16 conversations. Each conversation was approximately 2-3 minutes in length. The sampling rate of the motion capture system was 100 Hz. Figure 4.3 shows the position data from one trial on the x-y plane, with the participant's positions in orange, and the interlocutor's positions in blue, connected by arrows to show the path taken by the interlocutor.

To ensure systematic error was not present in the measurements, participants were instructed to reset their feet, shoulder, and head orientation to the same starting position between conversations.

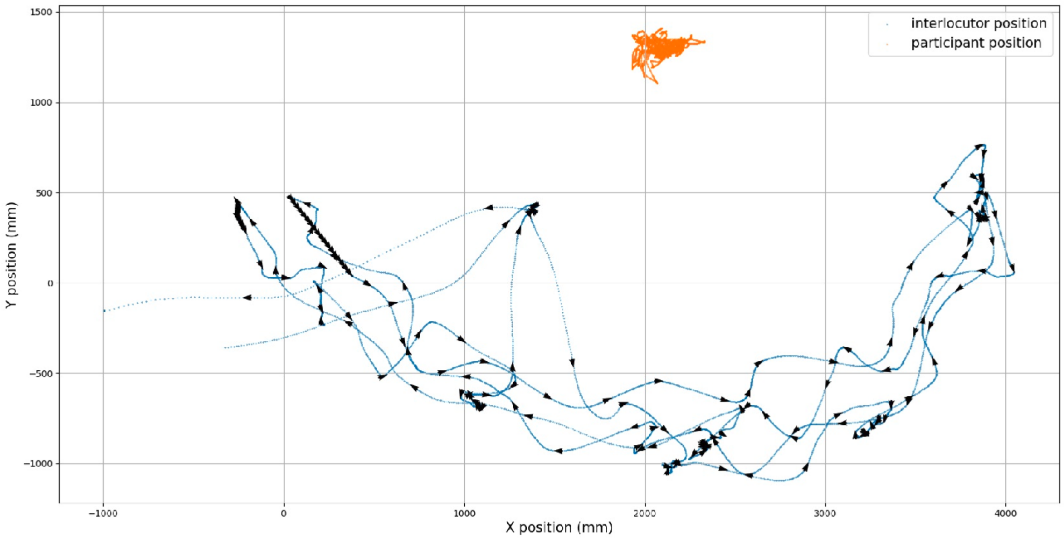

Figure 4.3: Example trial with interlocutor and participant positions

4.3 Data Analysis And Modelling

An extensive exploratory data analysis was carried out, to understand the ways in which data could be modeled and processed. Useful visualizations include the one shown in Figure 4.4, where the interlocutor angle and participant head and shoulder orientations are shown over time.

During conversation, participants will, on occasion, turn their head away from the interlocutor. Thus, the human data includes occasions where the behaviour is not tracking the interlocutor. Further, in addition to moving their feet, torso, and head, humans can also move their eyes to track an interlocutor. Thus, it was not expected that the head orientation would directly align with the direction of the interlocutor.

4.3.1 Convention

The convention used for the analysis is to model the three segments as cylinders that may rotate. Each segment has its own orientation, and is depicted in Figure 4.5. Figure 4.6 shows how the interlocutor angle is measured in the study space from a bird's eye view. We use the following variables to describe the speaker's configuration:

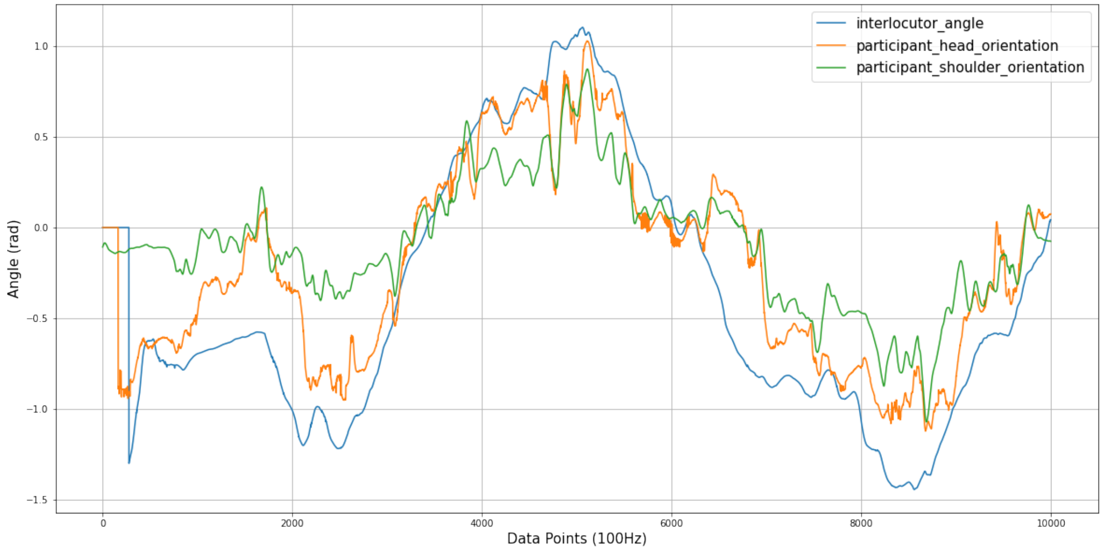

Figure 4.4: Example trial with interlocutor angle and participant head, shoulder orientation

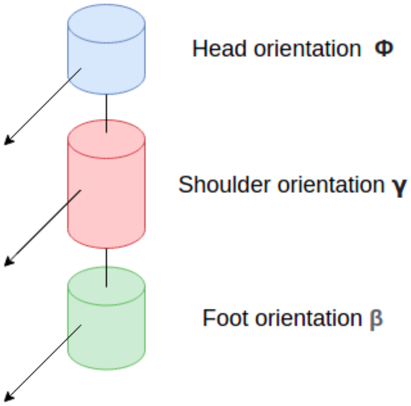

Figure 4.5: Representation of participant head, shoulder and foot orientation as connected cylinders

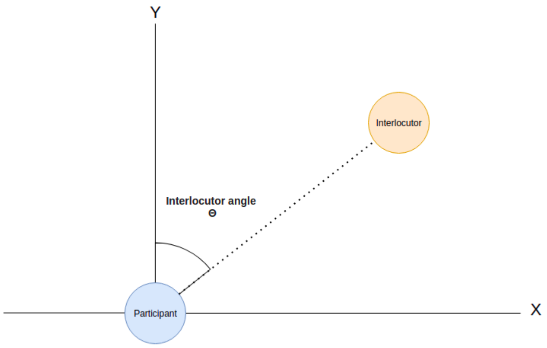

Figure 4.6: Representation of study space and interlocutor angle with respect to participant

$$\begin{aligned}\theta &= \text{interlocutor angle}\\ \phi &= \text{head orientation}\\ \gamma &= \text{shoulder orientation}\\ \beta &= \text{foot orientation}\end{aligned} \tag{4.1}$$

4.3.2 Formulation for Head and Shoulder Orientations

Inspired from [15], the head and shoulder models are coupled. Additional terms are introduced to also drive the head and shoulders with the interlocutor angle. These are given in (5.4) and (5.5).

$$\ddot{\phi} = k_1(\gamma - \phi) + k_2(\theta - \phi) + b_1(\dot{\gamma} - \dot{\phi}) + b_2(\dot{\theta} - \dot{\phi}) \tag{4.2}$$

$$\ddot{\gamma} = k_1(\phi - \gamma) + k_2(\theta - \gamma) + b_1(\dot{\phi} - \dot{\gamma}) + b_2(\dot{\theta} - \dot{\gamma}) \tag{4.3}$$

The use of both the angular displacement and the angular velocities results in a spring-damper style system. This acts as an extension based on the work in [15], where the head and shoulders are also modeled via a spring-damper system, and now include consideration for the interlocutor's angle. Given that the data is captured at the same rate of 100Hz, resulting in a constant time-step between all data points, the above formulations were discretized to allow for a direct fit of the data in terms of samples at different time steps.

Specifically, the backward difference and central difference formulations for the first and second derivatives will be applied to transform the written formulations. These approximations can be substituted into the original formulations. If used in (5.4), the result is as follows.

$$\begin{aligned}\frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{\Delta t^2} &= k_1(\gamma_i - \phi_i)\\ &+ k_2(\theta_i - \phi_i)\\ &+ b_1(\frac{\gamma_i - \gamma_{i-1}}{\Delta t} - \frac{\phi_i - \phi_{i-1}}{\Delta t})\\ &+ b_2(\frac{\theta_i - \theta_{i-1}}{\Delta t} - \frac{\phi_i - \phi_{i-1}}{\Delta t})\end{aligned} \tag{4.4}$$

Equation 4.4 now represents all the variables required in terms of the measured data at any given time-step. It also isolates the desired variable which is the next head orientation, ϕ_{i+1} in terms of all the other variables at the current or previous time-step. The isolation of the next head orientation simplifies to (4.5).

$$\begin{aligned}\phi_{i+1} &= k_1(\gamma_i - \phi_i)\Delta t^2 + k_2(\theta_i - \phi_i)\Delta t^2\\ &+ b_1(\gamma_i - \gamma_{i-1} - \phi_i + \phi_{i-1})\Delta t\\ &+ b_2(\theta_i - \theta_{i-1} - \phi_i + \phi_{i-1})\Delta t + 2\phi_i - \phi_{i-1}\end{aligned} \tag{4.5}$$

This formulation allows for fitting data at the current and previous time-steps for all variables in question, in order to estimate the head orientation at the next time-step. The same steps can be applied to the shoulder orientation formulation, and results in (4.6).

$$\begin{aligned}\gamma_{i+1} &= k_1(\phi_i - \gamma_i)\Delta t^2 + k_2(\theta_i - \gamma_i)\Delta t^2\\ &+ b_1(\phi_i + \phi_{i-1} - \gamma_i - \gamma_{i-1})\Delta t\\ &+ b_2(\theta_i - \theta_{i-1} - \gamma_i + \gamma_{i-1})\Delta t + 2\gamma_i - \gamma_{i-1}\end{aligned} \tag{4.6}$$

Using these simplified formulations, the data was fit to the model. The model was chosen to be driven at the same rate as the captured data at first, and so the value of Δt is set to 0.01.

In order to carry out the parameter estimation, a least squares optimization was applied. This function minimizes the sum of squares for a set of equations, and will receive formulations (4.5) and (4.6) simultaneously to find the best 4 parameters for each set of processed data. The analysis is carried out trial by trial, and the results are investigated to identify the quality of the fit and the formulations.

4.3.3 Formulation for Stepping and Feet Orientation

After examining the manner in which humans chose to step, it is evident that stepping occurred last, when the angular displacement of the interlocutor was large. For smaller angular differences, rotations of the participant's head and shoulders are sufficient to maintain the conversation. Therefore, this formulation takes into account how far the shoulders have rotated before stepping needs to be performed.

The formulation is discretely defined on every time step considering the difference between the shoulder and feet orientations at that time step.

$$\beta_{\mathrm{i+1}} = \begin{cases} \beta_{\mathrm{i}} & if\ |\gamma - \beta_{\mathrm{i}}| < \delta \\ \beta_{\mathrm{i}} + \Delta\beta & else \end{cases} \tag{4.7}$$

Here β denotes the average or center of the orientation of the left and right foot, averaging the stepping behavior of both that takes place briefly after each other. The parameter δ is identified by finding the best threshold between the shoulders and feet orientations that trigger a step in the participant. The step size $\Delta\beta$ itself is determined by identifying the average change in feet orientation when participants do take steps.

The behaviour of both feet was analyzed in order to find the typical step size in terms of angular displacement. This was done by first identifying large rapid changes in the foot orientation, visible in Figure 4.9. Once these periods are identified, the average change in orientation is calculated across all trials. This value, $\Delta\beta$, is found to be 0.21 rad for all steps in the study. A similar approach was taken to find the threshold needed for stepping to occur. At the points where these steps were identified, the difference between the shoulder orientation and the foot orientation was measured to find the thresholds at which participants start to step. Across the study, this average value for δ was 0.35 rad.

4.4 Results

Using the head and shoulder orientation formulation, the results of the least squares fit were computed across conversations and across subjects, shown in Table 4.1. This helps quantify the behaviour of each unique participant, as well as average behaviour across the entire study.

Each conversation is considered unique, and so to better represent the study results, the average parameters are found as in Table 5.4. Tracking results are generated by driving

Table 4.1: Best Fit Parameters For Each Conversation

Subject	Conversation	k_1	k_2	b_1	b_2
1	1	7914.7	-15600.2	-116.9	145.9
1	2	-37797.9	6307.9	319.0	-45.1
1	3	27093.7	14108.2	-333.7	-115.4
2	1	14677.5	-8054.7	-218.6	124.5
2	2	1616.44	10410.7	-78.7	-77.9
2	3	5061.3	-7076.0	-115.5	100.6
2	4	-5867.7	19385.8	39.5	-255.3
2	5	-14445.6	4129.1	69.9	7.9
3	1	-22829.3	12768.7	157.9	-86.3
3	2	-26750.5	-6773.4	191.1	120.6
3	3	2773.5	16365.9	-116.4	-86.2
4	1	-11712.6	-20090.7	62.2	211.0
4	2	20489.5	11282	-275.7	-69.0
4	3	1768.6	-10698.1	-116.9	205.4
5	1	721.3	11437.1	-76.3	-75.4
5	2	9018.7	-16732.2	-156.4	199.8

the model equations by the received interlocutor angles. This simulates how tracking may occur on a robot, since the models can only be applied using the received interlocutor angles and the previously generated estimates as per (4.5) and (4.6). Computations are carried out via ROS, also allowing easy migration of models to the actual robot. The interlocutor angles are fed into the model at 100Hz, the same rate at which the motion capture data are generated and processed.

Figure 4.7 demonstrates results when the interlocutor angle for subject 3, conversation 3 is passed in to the model. The green curve is the estimated head orientation, and demonstrates good tracking of the interlocutor angle in blue. A similar situation occurs with the shoulder orientation in Figure 4.8, as it shows a smooth estimate for what a shoulder angle should be given the interlocutor angle.

Figure 4.9 compares model estimated foot orientation to the measured participant foot orientation in the same way. The model's steps occur at similar points in time, and the change in angles indicate that the step sizes are also quite similar to the participant. Further evaluation is performed by examining how the modeled steps coincide with the participants' real steps. Steps are defined by rapid changes in measured foot angles, in

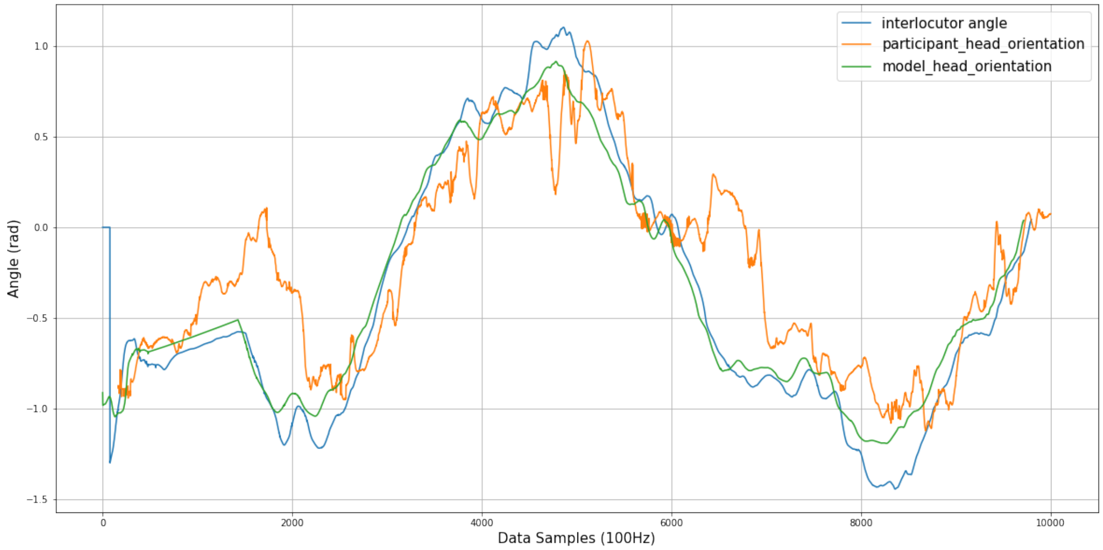

Figure 4.7: Example plot of interlocutor angle against measured head orientation and model estimated head orientation during a conversation

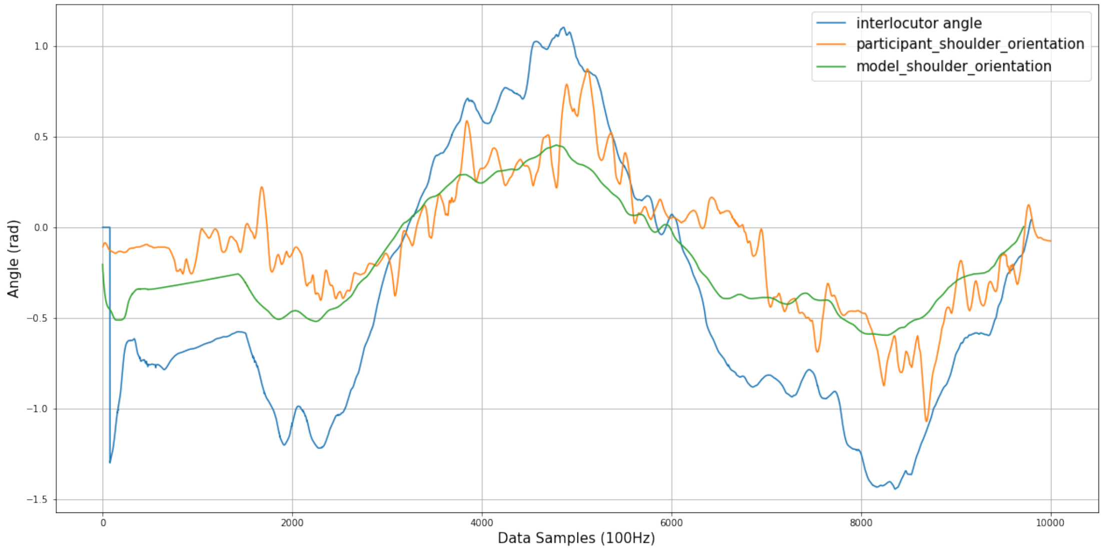

Figure 4.8: Example plot of interlocutor angle against measured shoulder orientation and model estimated shoulder orientation during a conversation

Table 4.2: Best Average Model Parameters

k_1	k_2	b_1	b_2
3189.8	3332.6	-79.7	11.2

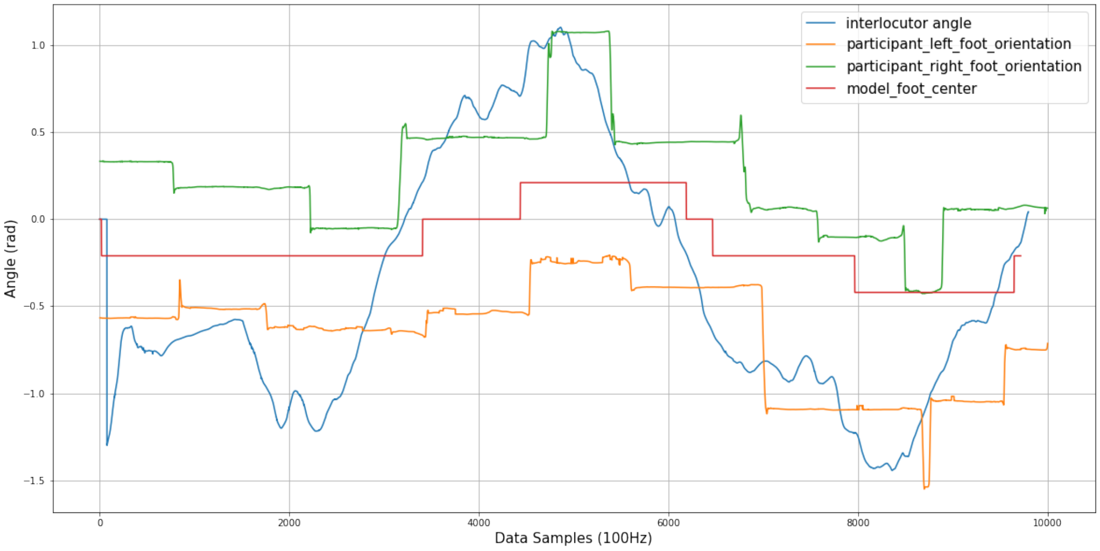

Figure 4.9: Example plot of interlocutor angle against measured foot orientation and model estimated foot orientation during a conversation

both the participants' and modeled steps. Each modeled step is considered accurate if it occurs within 1 second of a participant's observed step. An accuracy is generated on every trial, comparing the model's steps and steps made by both of the participant's feet, as shown in Table 4.3. It can be seen that for the majority of trials, stepping accuracy as a percentage is considerably high for at least one foot.

An error analysis was also performed by conversation, to highlight model performance quantitatively with the best fit parameters. Figure 4.10 and Figure 4.11 show squared residuals in histogram form for head and shoulder orientations, relative to the interlocutor angle, for both the participant and model.

The mean squared error (MSE) for the head and shoulder estimates were similarly recorded for all trials, with both the best fit and the average parameters, and shown in Table 4.4.

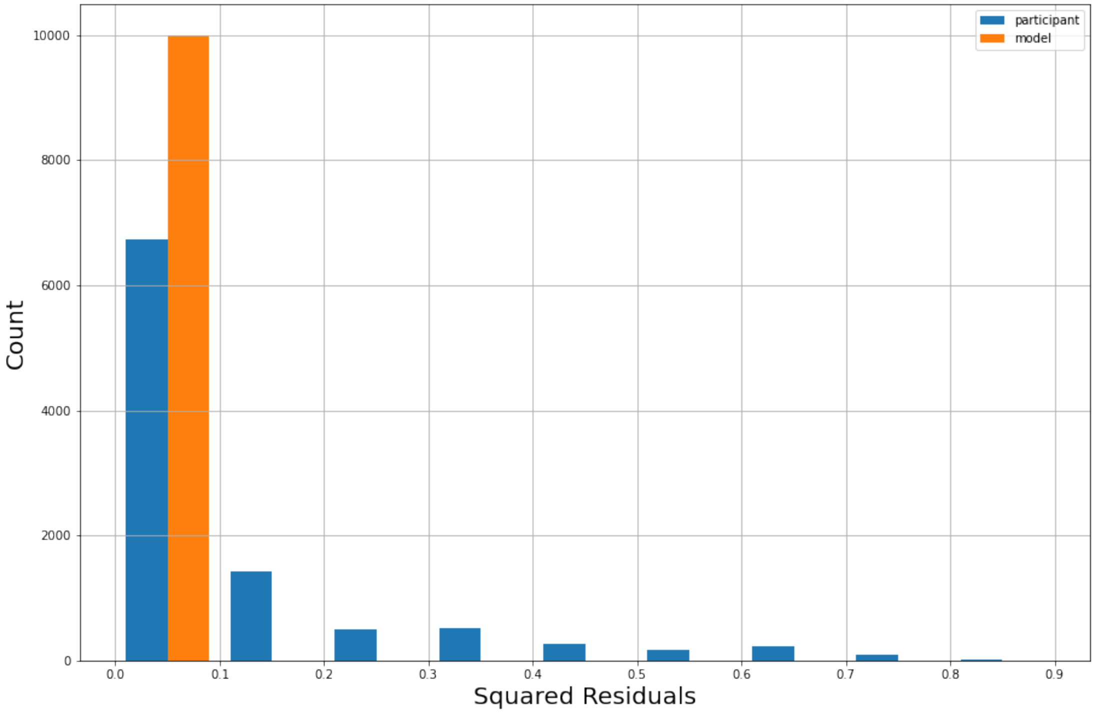

Figure 4.10: Histograms of differences between head and interlocutor angle for participants (blue) and model (orange) during conversation

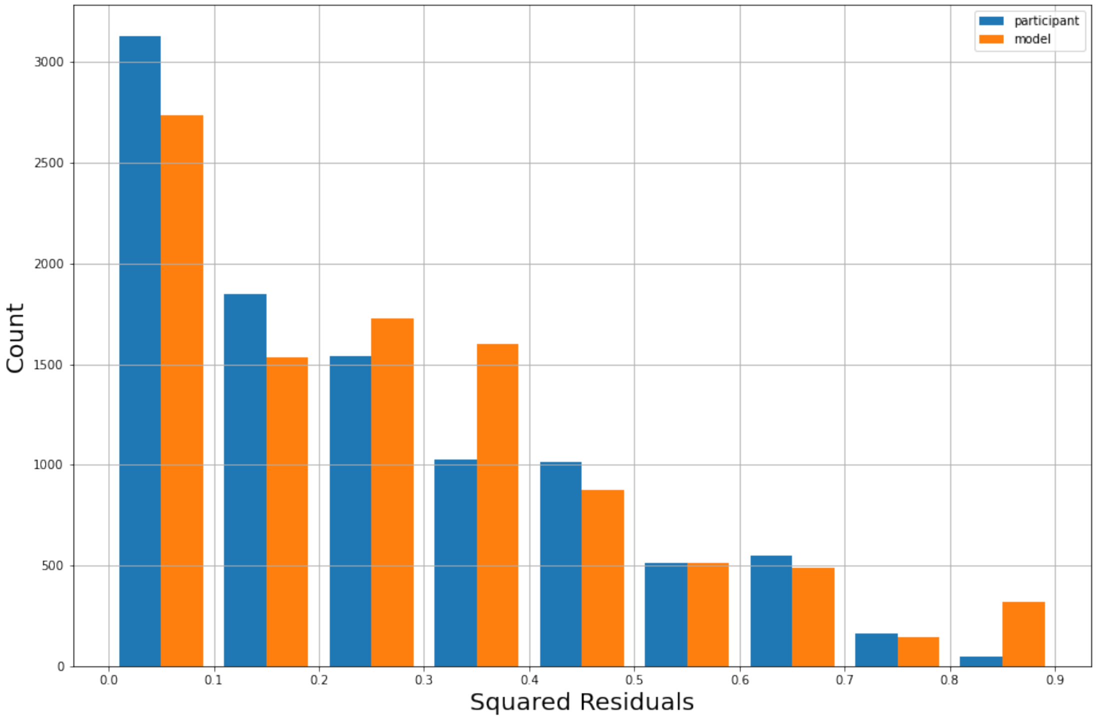

Figure 4.11: Histograms of differences between shoulder and interlocutor angle for participants (blue) and model (orange) during conversation

Table 4.3: Stepping Model Accuracy In % By Foot

Subject	Conversation	Left Foot Accuracy	Right Foot Accuracy
1	1	82	75
1	2	100	100
1	3	77	66
2	1	69	79
2	2	80	100
2	3	52	52
2	4	95	87
2	5	86	86
3	1	81	85
3	2	100	70
3	3	86	100
4	1	100	100
4	2	83	83
4	3	50	67
5	1	74	100
5	2	67	100

4.5 Discussion

For this study, analysis of the head, shoulder, and feet angle was performed for each participant. The data for each participant was aggregated across conversations and used to fit the model with the goal of trying to capture each individual's own unique behaviour. However, it was evident that generating good fits for the data was difficult when data was aggregated across conversations. This is due to two main reasons. First, each participant did not behave the exact same way across conversations. For instance, it was observed that some participants were reluctant to turn their shoulders in some conversations but not in others. As well, during conversations, it is natural for participants to look away from the interlocutor on occasion. The frequency and timing of this behaviour is not necessarily consistent across conversations. In addition, the conversations themselves are unique, in that the interlocutor does not move in the exact same way each time (although the general movement is similar across conversations). As a result, aggregated data does not generate good model fits.

When comparing results, it can be seen in some cases, lower MSE was obtained with av-

Table 4.4: Head And Shoulder MSE Per Conversation With Individually and Average Fitted Parameters

		Individual		Average	
Subject	Conversation	Head	Shoulder	Head	Shoulder
1	1	0.33	0.39	0.57	0.80
1	2	0.13	0.008	0.10	0.05
1	3	0.23	0.16	0.26	0.14
2	1	0.23	0.09	0.23	0.06
2	2	0.07	0.02	0.10	0.04
2	3	0.15	0.08	0.18	0.08
2	4	0.49	0.67	0.38	0.52
2	5	0.45	0.63	0.49	0.65
3	1	0.29	0.14	0.59	0.65
3	2	0.12	0.08	0.08	0.04
3	3	0.08	0.04	0.08	0.06
4	1	0.25	0.70	0.09	0.59
4	2	0.28	0.17	0.21	0.07
4	3	0.16	0.11	0.07	0.09
5	1	0.28	0.13	0.06	0.23
5	2	0.24	0.11	0.14	0.08

eraged model parameters rather than the individual fit for that conversation. This is likely due to the participant engaging in more instances of looking away from the interlocutor during that conversation.

A comparison of the interlocutor angle with the model and participant head angle, as shown in Figure 4.10, indicates that the model can track the interlocutor very well, with an MSE of 0.03. The participant tracks with an MSE of 0.11, and has a wider distribution of errors in the histogram. This is consistent with the hypothesis that humans use eye-gaze in addition to rotating their head and shoulders to track an interlocutor. The incidences of large differences between head angle and interlocutor angle are consistent with participants engaging in instances of looking away from the interlocutor. Nevertheless, tracking the average behaviour with the present model was successful. Figure 4.11 shows that model and participant shoulder behaviour are quite similar, and no significant discrepancies are seen. This suggests that the adding look-away behaviour to a model only needs to incorporate head motion.

In Table 4.3, the accuracy of the stepping model is not always the same for both feet. Further investigation shows this is because participants do not always perform steps with both feet to naturally reorient themselves to their conversational partner. Trials with low accuracy are explained by recognizing that participants do not always perform steps even if their shoulders are rotated past a certain threshold. This appears to be mostly up to the participant's discretion, and explains why the accuracy is low for subject 2, conversation 3, but is high for subject 2, conversation 2. Future work could look to characterize this behaviour with greater precision.

4.6 REEM-C Implementation Of Speaker Tracking

This work was implemented on the REEM-C Humanoid Robot to simulate a real human-robot interaction scenario. The algorithms developed in the previous two sections may be used to estimate the angle of the interlocutor as they move around in front of the robot. This estimation and usage is described in further detail in the next chapter, regarding the fusion and integration and audio and video information.

Stepping is generated on both feet separately, with a focus on the direction that the step needs to take. As per the study data, as the interlocutor moves to the participant's left, the left foot steps first followed by the right foot, and vice versa. This functionality is also present on the robot.

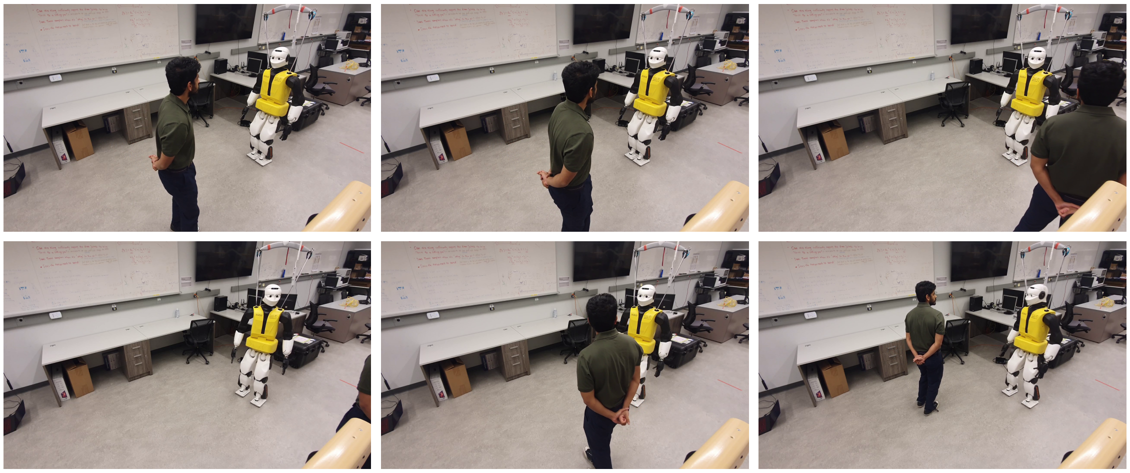

Figure 4.12: Frames depicting the REEM-C moving its head, shoulders and feet to orient to a moving conversational partner (view from top left to bottom right)

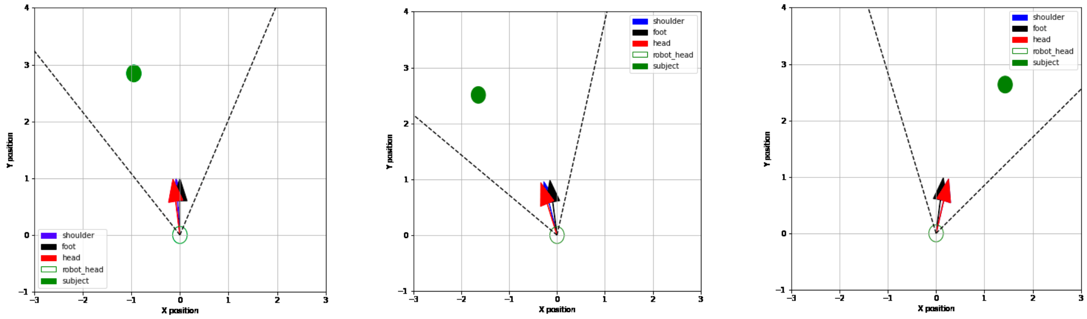

Figure 4.13: Bird's eye animation of REEM-C, along with its field of view shown in black dotted lines, depicting a visually detected conversational partner shown as the circle. The circle is filled with a red dot when the audio DOA is closely associated with the visual DOA of the subject

Figure 4.13 demonstrates the REEM-C in different head, shoulder and foot orientations as it tracks a moving interlocutor.

4.7 Conclusions

In this chapter, we present a model for the head, body and foot orientation in the context of tracking moving conversational partners. This model and its implementation allow for realistic, human-like tracking of an interlocutor by a humanoid robot.

Extensions for the work could include incorporating look-away behaviour that participants can exhibit during natural conversations. Participants also behave differently when speaking compared to when listening, and so different models and approaches can be designed to account for the difference in behaviour to make human-robot interaction even more natural.

Chapter 5

An Audio-Video Sensor Fusion Framework To Augment Humanoid Capabilities For Identifying And Interacting With Human Conversational Partners

5.1 Introduction

Human-robot interaction (HRI) is a necessary field to study because it enables humans to communicate with and control their robotic companions, making it possible to integrate them into various aspects of human life and work. This can lead to improved efficiency, safety and expanded capabilities in many different industries and fields. Works in human-robot interaction often involve improving the communication loop between a single human and a robot, or have a direct focus on generating natural speech patterns and non-verbal behaviours to allow humanoids to integrate socially. [1] combines speech recognition and gaze to allow robots to assist humans in assembly tasks, in primarily stationary scenarios. [6] and [19] explore the emotional affect of humans and robots to make HRI more natural. [1] Works such as [31] and [39] use audio intelligence to identify many potential sound sources for robotics applications. Other works such as [46], [28], [27] and [8] use both

[1] The content of this chapter is from the following conference paper: Pranav Barot, Ewen MacDonald, Katja Mombaur, An Audio-Video Sensor Fusion Framework To Augment Humanoid Capabilities For

audio and video to help robots identify active speakers for HRI, but do not present a framework for responding to what they are able to detect. [34] performs particle filtering of the measured audio and video data, but does not use their approach in real time for a real HRI application.

There is a need to improve humanoid interactions with potentially multiple moving human subjects in a wide-ranged workspace. These environments can involve a humanoid robot and several human subjects, some of whom may be speaking and some who may not. It is desirable for the humanoid to intelligently navigate a complex environment by tracking moving human subjects, correctly directing attention to humans who may require it and facilitating interactions with multiple humans at once. Our work addresses this problem by introducing a multi-modal framework to handle more dynamic HRI scenarios. The presented work helps accommodate for noisy and busy environments, while also considering the most realistic and appropriate behaviours for humanoid robots to exhibit. This work is needed to help accelerate the integration of robots into society, where they may be required to work alongside or take care of humans. The direct applications for this work extend to social robotics, surveillance or human-robot collaboration in a variety of workspaces such as warehouses or labs. The potential for further humanoid capability is opened up with this research, as humanoids may then augment their human interactions with manipulation, gesturing or locomotion.

5.1.1 Contributions

Our work distinguishes itself by developing a complete sensor fusion framework using audio and video intelligence that is directed towards HRI, performing a full systems integration, and utilizing and evaluating the system for real-time usage. The work helps fill the gaps where individually developed systems may be used in tandem to account for their shortcomings, and augment the intelligence of humanoids for HRI purposes.

The contributions for this work are as follows.

- Implementation and deployment of robust, online audio and video intelligence systems onto the REEM-C
- A sensor fusion framework that integrates all sub-systems and generates natural motions on the REEM-C

Identifying And Interacting With Human Conversational Partners, In Conference on Intelligent Robots and Systems (IROS), Detroit, USA, 2023

- Quantitative validation of developed approaches on recorded data in a lab environment
- A user study with evaluations of overall robot intelligence and behaviour using designed approaches

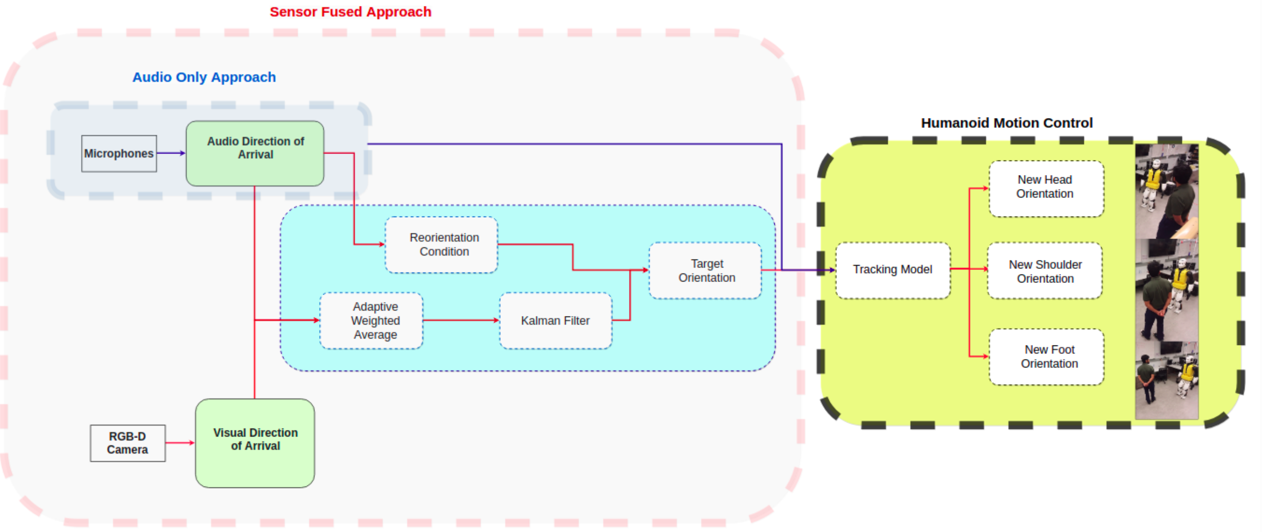

Figure 5.1: Flowchart for proposed work. The audio only approach is highlighted with the light blue border. The sensor fused approach is highlighted with the red border.

The overall process for this proposed work is visualized in Figure 5.1. Section 5.2 explains the main algorithms and hardware used for audio and video intelligence. Section 5.3 highlights ways in which these algorithms are fused. Section 5.4 outlines the design of additional acoustic intelligence on the REEM-C, whereas Sections 5.5 and 5.6 discuss the behavioural model and usage of the REEM-C in the described scenarios. Section 5.7 outlines the protocol and results for evaluating the robot's behaviour on real human subjects, and Sections 5.8 and 5.9 provide a discussion on the results of the proposed approaches and key conclusions.

5.2 Algorithmic Framework

The humanoid used for this work is the REEM-C made by PAL Robotics, which we control via ROS. The REEM-C comes with an RGB-D camera in the forehead, but no microphones. We use two Comica microphones, connected to a Scarlett 2i2 audio interface for the purposes of the audio intelligence used in this work. A taut headband is placed on the REEM-C's head, allowing easy installation and removal of the microphones. The cost of the additional hardware is no more than $350. This is a generalizable setup and costs are minimal, allowing this work to be replicated and used on all robots. We use an algorithm that integrates a number of key SSL concepts to perform accurate direction of arrival estimation of human speech using interaural timing differences (ITD) with two microphones on the REEM-C Humanoid Robot. The specifics of the technique, optimizations and real-time considerations are outlined in our work here [29], which is currently under review.

We choose this approach to the audio intelligence over other solutions such as the Amazon Alexa, for interfacing and performance reasons. The data processed on these prebuilt devices is not easily accessible, and testing suggests the accuracy of DOA estimation is not reliable in reverberant environments. Using a binaural microphone setup allows for a realistic deployment on the robot as it resembles ear placement on humans. It is also a worthwhile research interest to implement a binaural DOA estimation pipeline, as opposed to using arrays with several microphones, as described in [29].

Video systems are a series of algorithms used to estimate the behaviour of interlocutors for the purposes of human-robot interaction [3]. These algorithms are used in conjunction on the REEM-C, and lay the foundation for more novel and complete HRI frameworks. The visual estimation of the angular displacement of interlocutors is of primary focus, as this information coincides with audio SSL and provides the potential for a relevant sensor fusion approach. This video information is generated via the RGB-D camera present in the REEM-C's head.

Our aim is not to use state-of-the-art hardware and algorithms to optimize detection and accuracy, but to use a minimal set of easily developed and deployed sub-systems that can be used on the REEM-C for effective HRI experimentation. This is why we aim to use a simple binaural microphone setup and only the camera present in the REEM-C's head, as opposed to more expensive and performant alternatives.

5.3 Sensor Fusion Approaches

The described algorithms are used concurrently in an online fashion, and need to be fused to allow for a more complete interaction framework. A few approaches are explored to combine the video and audio information.

The generated information, as well as the current state of the robot, is recorded and synthesized via communication across different ROS nodes. We ensure that received information is for the same timestamp considering the different sampling rates of each individual process, allowing for accurate and responsive robot behaviour to occur in real time. The synchronized behaviour will be limited by the slowest occurring process, which is the audio direction estimation, running at approximately 6 Hz. We explore a number of methods to fuse this information for the purposes of accurately responding to the environment of the robot.

5.3.1 Choosing the active speaker

In a scenario with multiple subjects present, the robot requires a method to decide which subject to follow. This is decided by checking the agreement of audio and video direction of arrival, via the following

$$detected\,speaker = \operatorname*{argmin}_i(video_{DOA_i} - audio_{DOA}) \tag{5.1}$$

where $video_{DOA}$ is the vector containing all detected visual DOA. The detected speaker's overall DOA is then used to facilitate the tracking behaviour of the robot, which is designed to alternate between speakers as and when they are identified. Figure **??** depicts a trial with video and audio DOA recorded for a single participant as they move around in front of the robot. Errors in the video DOA may come from unsteady facial landmark detection in motion, unreliable choice of target points, or scaling issues with depth images recorded on the robot. Errors in the audio DOA come from reflections, other noises in the environment and the overall difficulty of online audio processing. Therefore, both forms of information will need to be appropriately synthesized to ensure robot behaviour is accurate.

5.3.2 Tracking behaviour: fixed weight estimation

One method to combine received information for the robot to track is an inverse variance weighted average, with weights reflecting the confidence in the measurements of each specific sensor.

$$r_{DOA_{fixed}} = \frac{\sigma_{video}^{-2} * video_{DOA} + \sigma_{audio}^{-2} * audio_{DOA}}{\sigma_{video}^{-2} + \sigma_{audio}^{-2}} \tag{5.2}$$

Sensitivities in the sensors are accounted for by measuring the variance of estimates at different angular displacements relative to the robot. Data is recorded at 15 degree intervals, and is considered to account for ± 7.5 degrees. Video and audio estimates are measured in the robot lab environment at each interval, and the variances are recorded in Table 5.1. The variances are generalized to describe the behaviour on both sides of the robot.

Table 5.1: Variance and calculated weights for direction estimates from both modalities. Video estimates for angles greater than 30 degrees are not available as they are beyond the field of view of the camera used

ANGLE (deg) ± 7.5	VIDEO VARIANCE	AUDIO VARIANCE	VIDEO WEIGHT	AUDIO WEIGHT
75	N/A	0.0640	N/A	16.0
60	N/A	0.0044	N/A	230.0
45	N/A	0.0039	N/A	260.0
30	0.0028	0.0031	360.0	320.0
15	0.0010	0.0027	960.0	370.0
0	0.00051	0.0011	1950.0	910.0

Video estimates are also affected by the performance of the facial landmark detector, which fails to detect subjects at a distance of 3.1m or farther from the robot. The weights are normalized on each interval for both the video and audio systems and visualized in Figures 5.2 and 5.3.

Unsurprisingly, the variance for both methods is lowest when subjects are directly in front of the robot, and increases as subjects move towards larger angular displacements. We believe that the variance of audio estimates past 60 degrees is much higher than in the other ranges due to much more significant occlusion of sound waves by the robot head in between the microphones.

5.3.3 Tracking behaviour: adaptive weight estimation

Another method to resolve the received information involves adapting the weights to the behaviour of the system. This results in a recursive adaptation of the Fraser-Potter smoother [11], which performs sensor fusion by adhering to the reliability of the sensors online. This is shown in Figure 5.3, where rather than using inverse variances as weights, the sensors

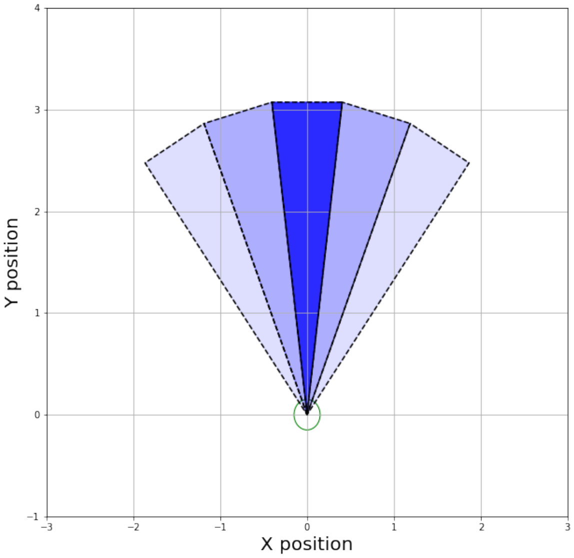

Figure 5.2: Normalized weights for video DOA estimation, limited to robot FOV and distance of 3.1m. Robot head shown at (0,0)

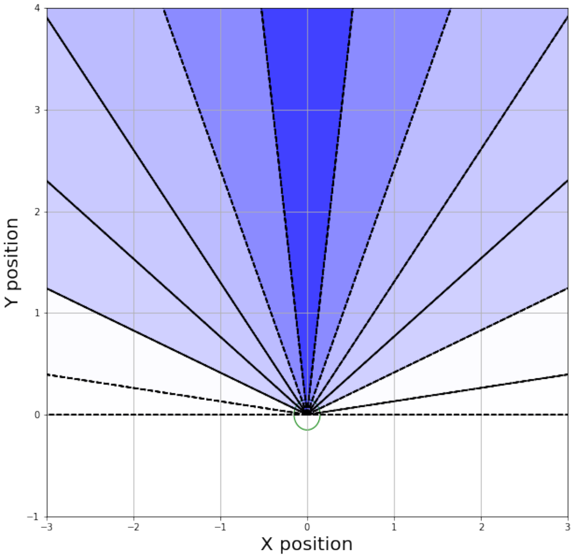

Figure 5.3: Normalized weights for audio DOA estimation, limited to 180 deg range. Robot head shown at (0,0)

are scaled by the variance of their complementary measurements.

$$r_{DOA_{adaptive}} = \frac{\sigma^2_{video} * audio_{DOA} + \sigma^2_{audio} * video_{DOA}}{\sigma^2_{video} + \sigma^2_{audio}} \tag{5.3}$$

This method will estimate the variances of the video and audio information using the past N estimates for both sensors. If video estimates happen to falter due to occlusion or rapid movement, or audio estimates are influenced by reflections and activity around the robot, their respective weights will be adjusted to reflect this behaviour accordingly. The expected results with a good choice of N should produce a steadier trajectory than with fixed weights.

The choice of N comes from inspecting collected data and generating trajectories by adjusting the value of N. A good value of N should be large enough to capture the system's dynamics, while not being too large so as to negatively influence the behaviour at the current time. The best value for N is determined qualitatively by inspecting the smoothness of the generated trajectory, and quantitatively by checking how close the trajectory is to the audio and video estimates. The results are checked for values of N ranging from 5 to 25, and the mean squared error between the generated result and the video and audio estimates are recorded separately. We aim to choose the value of N which minimizes the joint product of the video and audio mean squared errors. This joint product is shown in Figure 5.4 for an example trajectory.

Across similar trials, it appears the value of N = 8 is a reasonable value to maintain sufficiently smooth estimates and not require the system to hold a large amount of memory. The result of the fixed weighting against the adaptive weighting for N = 8 is shown in Figure 5.5.

This result indicates the adaptive weighting is more reliable overall and generates smoother trajectories compared to the fixed weighted approach.

5.3.4 Tracking behaviour: Kalman Filter

In order to send natural, realistic trajectories to the robot, it is important to filter out abrupt changes in the measurements made. We employ a Kalman filter for the fused audio and video information to address this problem. We believe that a Kalman filter used on the fused information is a better solution than just a Kalman filter on both measurements, as the latter approach would still require a measurement transition matrix that best combines the filtered audio and video information. With the proposed adaptive weighted average, it would be more advantageous to use the Kalman filter after this step.

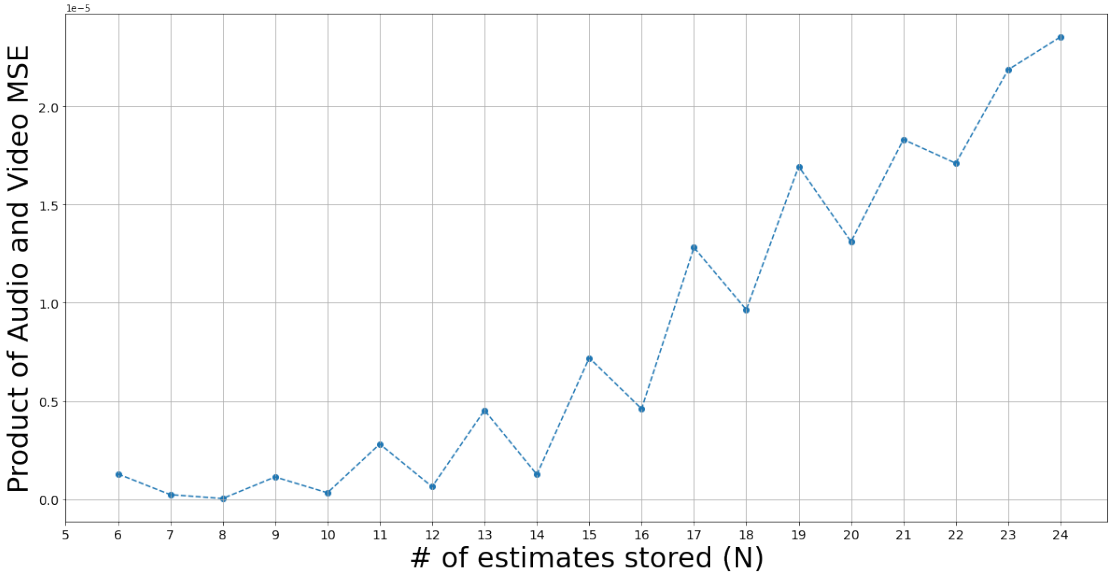

Figure 5.4: Product of MSE to video and audio estimates as a function of lookback size N for variance estimation

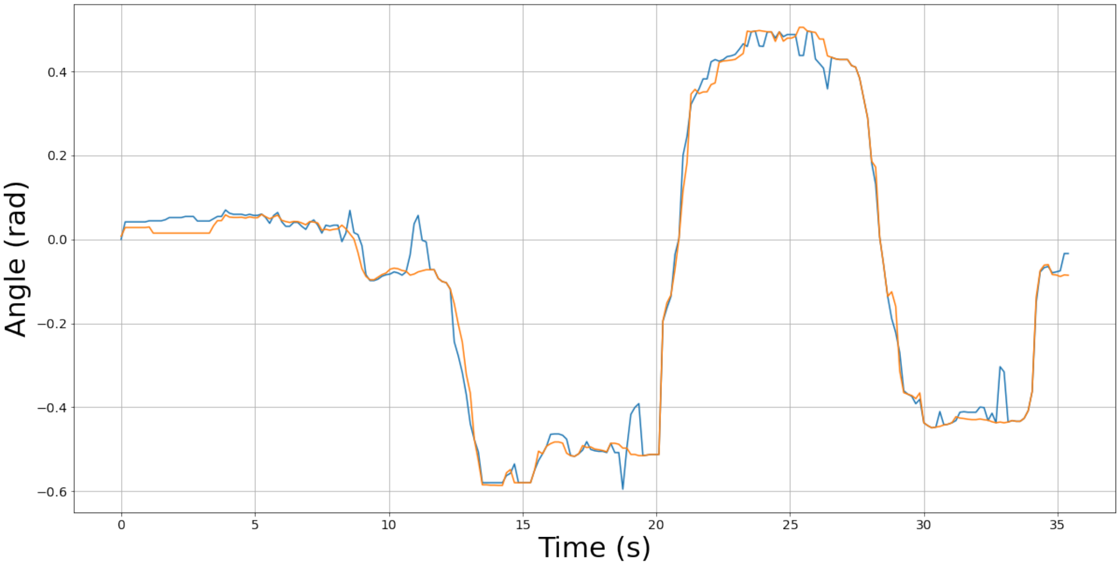

Figure 5.5: Visualizing result of fixed weighted average (blue) vs. adaptive weighted average (orange) with N = 8

The state matrix of the filter will be the position and velocity of the measured signal. The measurement variance R is directly computed from the adaptive weighted average of trials with subjects moving around and speaking to the robot.

$$R = \begin{bmatrix} 0.116 \end{bmatrix}$$

The initial state covariance matrix P_0 comes from analyzing the recorded data and computing the variance of the angular position and velocity. Since the measured position ideally should not be correlated to the velocity, the off-diagonal is set to 0.

$$P_0 = \begin{bmatrix} 0.116 & 0 \\ 0 & 0.046 \end{bmatrix}$$

The process noise covariance matrix assumes some further uncertainty in the state variables as a result of variations across time, and so factors in the timestep as follows in Q.

$$Q = \begin{bmatrix} 0.116 \cdot \Delta t^2 & \frac{0.116 \cdot \Delta t^3}{2} \\ \frac{0.116 \cdot \Delta t^3}{2} & 0.116 \cdot \Delta t^2 \end{bmatrix}$$

This information is used to process the averaged DOA and generate new estimates for the fused sensor output using the standard Kalman filtering steps.

$$\hat{x}_k = A\hat{x}_{k-1}$$

$$P_k = AP_{k-1}A^T + Q$$

$$K_k = P_k H^T (HP_k H^T + R)^{-1}$$

Using the Kalman gain K_k, the update steps then take place.

$$\hat{x}_k = \hat{x}_k + K_k(z_k - H\hat{x}_k)$$
$$P_k = (I - K_k H)P_k$$

The Kalman filter is applied to measured data after performing the adaptive weighted average, and generates a result as in Figure 5.6.

The red circles indicate points where the adaptive weighted average would result in abrupt changes in the DOA estimates. The real life implication of this would result in the REEM-C performing noticeably sudden readjustments that may take away from the smoothness and realism of the interaction. The Kalman filter handles abrupt changes in the measurement by providing a smoother transition through variable estimates, which will result in more natural robot behaviour.

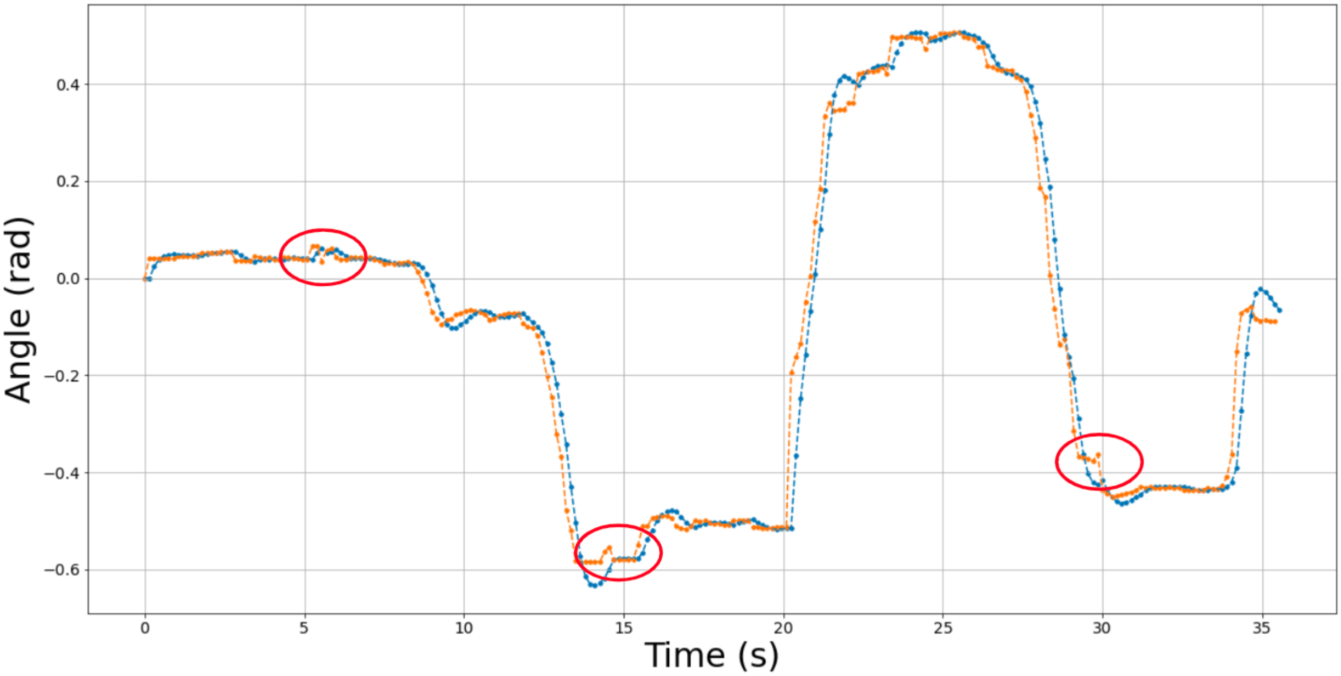

Figure 5.6: Result of Kalman filtered output (blue) plotted against adaptive weighted average (orange). Red circles indicate key differences in both outputs.

5.4 Reorientation Study

Determining when the robot should look away and check for other subjects outside the field of view of the camera is not as simple as directly following the measured audio DOA. A decision must be made about when to continue the current tracking behaviour and when to look away.

A study was performed to evaluate this scenario in greater detail. Trials were recorded with subjects speaking while they were positioned outside the field of view of the REEM-C. For the purposes of being able to identify when sounds should be attended or should be ignored, the participants also generated non-speech sounds such as coughs or shuffles of nearby chairs. This simulates scenes in a workspace where tools are being used, or doors are being opened or closed, and should therefore not require a robot's attention. An example trial is demonstrated in Figure 5.7.

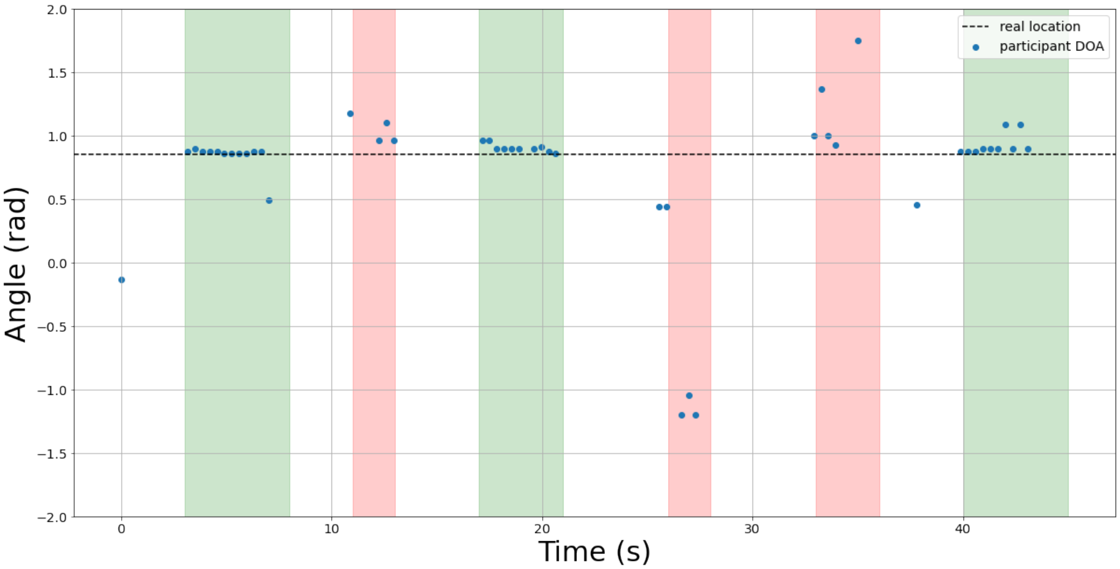

Figure 5.7: Example trial with SSL results as blue dots. The green regions indicate intervals with speech. The red regions indicate intervals with non-speech sounds.

A total of 6 such trials were performed, with one or two subjects participating in each trial. It can be seen that for speech sounds in the periods of green, the direction of arrival is very steady, but for artificial sounds, the estimates are far less so and appear to come from a variety of different locations. A similar trend is seen in all other trials. We aim to use this property to classify whether or not reorientation is required.

A sliding window approach is used on the direction of arrival estimates to evaluate this phenomenon with a window size of 5 and a step size of 1. A number of metrics are evaluated against the annotated trials to determine the best way to identify the need for reorientation. Since legitimate sounds generate steadier estimates of DOA, an ideal distribution is constructed as being the average of the estimates in each window. This ideal distribution is compared to actual received estimates using the cosine similarity, the KL-divergence and Wasserstein distance. These metrics should quantify how different the observed distribution is from the ideal distribution. In order to evaluate steadiness, the range and standard deviation of the measured estimates is also taken into account, by thresholding the product of these two metrics. Thresholds are empirically generated for each of the proposed metrics by inspecting the behaviour across all 6 trials. An initial accuracy is computed with each of these methods in Table 5.2.

Table 5.2: Accuracy For Reorientation Trials By Method

METHOD	TRIAL 1	TRIAL 2	TRIAL 3	TRIAL 4	TRIAL 5	TRIAL 6
Cosine Similarity	97.6	89.6	83.1	66.7	76.7	62.1
Range-Std Method	83.9	91.4	81.3	70.2	60.4	64.6
KL Divergence	63.5	58.4	65.4	65.4	67.7	46.7
Wasserstein Distance	61.9	46.7	54.9	66	70.1	66.9

Further classification results are presented for the trials with the cosine similarity in Table 5.3.

Table 5.3: Classification Results For Trials With Cosine Similarity

METRIC	TRIAL 1	TRIAL 2	TRIAL 3	TRIAL 4	TRIAL 5	TRIAL 6
Accuracy	97.6	89.6	83.1	66.7	76.7	62.1
Precision	97.1	95.7	88.9	89.1	94.1	77.9
Recall	94.3	75.9	60.0	47.1	71.9	66.7
F1 Score	95.7	84.6	71.6	61.7	81.5	71.9

5.5 Robot Tracking And Behavioural Model

The behaviour of the robot is determined by a mathematical model that controls the rotation of the head, shoulders and feet in response to measured interlocutor angles. Further details of this model are available here [2], which is currently under review.

Below are the formulations used on the robot and the required parameters for the robot head orientation ϕ, the shoulder orientation γ and the orientation of the feet β given a measured interlocutor angle θ.

$$\ddot{\phi} = k_1(\gamma - \phi) + k_2(\theta - \phi) + b_1(\dot{\gamma} - \dot{\phi}) + b_2(\dot{\theta} - \dot{\phi}) \tag{5.4}$$

$$\ddot{\gamma} = k_1(\phi - \gamma) + k_2(\theta - \gamma) + b_1(\dot{\phi} - \dot{\gamma}) + b_2(\dot{\theta} - \dot{\gamma}) \tag{5.5}$$

$$\beta_{\mathrm{i+1}} = \begin{cases} \beta_{\mathrm{i}} & if\ |\gamma - \beta_{\mathrm{i}}| < \delta \\ \beta_{\mathrm{i}} + \Delta\beta & else \end{cases} \tag{5.6}$$

Table 5.4: Tracking Model Paramters (From [2])

k_1	k_2	b_1	b_2	δ	$\Delta\beta$
3189.8	3332.6	-79.7	11.2	0.35	0.21

With the Kalman-filtered result, the behaviour of the head and shoulders are depicted in Figure 5.8.

Noticeably, the head moves first to follow the interlocutor, and the shoulders follow soon after. This creates a very natural looking, realistic behaviour on the humanoid robot that mimics what humans perform in similar scenarios.

While a subject is being tracked using the filtered DOA estimates, the reorientation condition is checked with the collected audio DOA measurements. If the condition is triggered, the new target shifts from the moving subject to the angle that triggered the reorientation condition, as shown in 5.1. Given a filtered DOA estimate from the tracking process, the target angle changes if the reorientation condition is triggered.

$$similarity = \frac{measured \cdot ideal}{\|measured\| \, \|ideal\|} \tag{5.7}$$

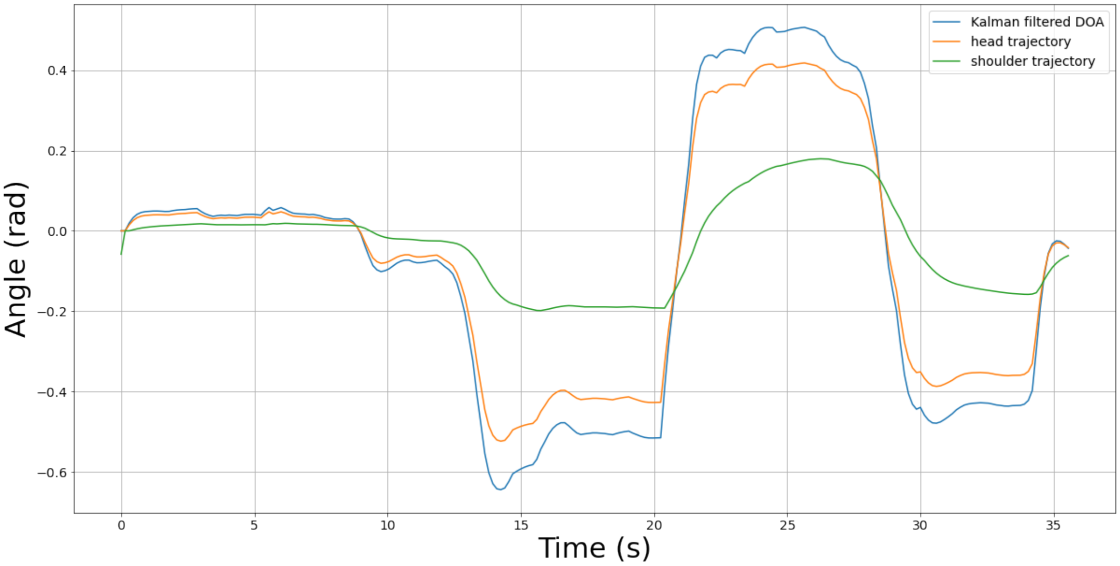

Figure 5.8: Head and shoulder trajectories on REEM-C given a filtered interlocutor angle, all in an absolute frame

$$external\ angle = \frac{1}{5}\sum_{i=1}^{5} measured_i \tag{5.8}$$

$$Reorientation = \begin{cases} external\ angle & if\, similarity > 0.85 \\ Filtered\ DOA & else \end{cases} \tag{5.9}$$

5.6 Robot Deployment And Online Usage

The REEM-C is a humanoid robot, with the robot's base frame present in the pelvis. In this experiment, the robot primarily undergoes yaw motions for each segment, to allow for following human subjects as they walk around in a large space. Pitch motions may be incorporated in the future to allow for more realistic behaviour, and to allow the REEM-C to interact with elements of the environment below its line of sight.

The motions of the REEM-C joints occur on their local frame. Thus, given a unique head yaw angle ϕ and a shoulder yaw angle γ, the resultant orientation of the robot's head with respect to the base frame will be

$$\epsilon = \phi + \gamma \tag{5.10}$$

With a stepping angle of β, the resultant orientation of the robot's head in the base frame remains ϵ as stepping results in the base frame's rotation as well.

This requires a resolution, as interlocutor information is received with respect to the head's orientation, as that is where the camera and microphones are mounted. If a measured interlocutor angle θ_{rel} is received, then their angle in the base frame's coordinate system is

$$\theta_{base} = \epsilon + \theta_{rel} \tag{5.11}$$

As a result, the online deployment requires using the ROS *joint states* information to estimate ϵ, and then sends the correctly filtered value of θ_{base} to the tracking model. Reliable synchronization between the ROS nodes running at different rates is key to ensuring the behaviour of the robot is accurately reflective of the scene at every instant. We utilize carefully designed ROS topics that ensure timing delays from communication and computation are negligible and do not affect real robot behaviour.

5.7 Results And User Study

The proposed framework is deployed onto the REEM-C with installed microphones. A quantitative evaluation is carried out using the designed approaches and real measured data. The Vicon motion capture system was used to measure the ground truth angular displacements of a moving participant with respect to the REEM-C. This data, alongside synchronized audio and video DOA estimates, were used to evaluate the accuracy of the presented approaches. A total of 4 trajectories are collected and evaluated using the fused approach with only the adaptive weighted average, and the fused approach using the adaptive weighted average and the Kalman filter. Figure 5.9 shows the video/audio estimates, alongside the result of the proposed sensor fusion approaches.

The average mean squared error is found to be 0.050 for the adaptive weighted average technique, and 0.042 when the Kalman filter is added. This difference in accuracy is not considered significant, but as previously described, the smoothing effects of the Kalman filter make it more viable for an HRI application.

We also performed a small user study with 5 participants for each approach, and recorded their evaluation for each trial in terms of naturalness, accuracy, responsiveness, and an overall rating out of 5. Subjects were instructed to move wherever they choose and speak to the robot as if they were having a conversation. Each trial lasts about 2-3 minutes, and the procedure for this study was approved by the University of Waterloo Research Ethics Board. We aim to evaluate behaviour with only audio tracking compared to

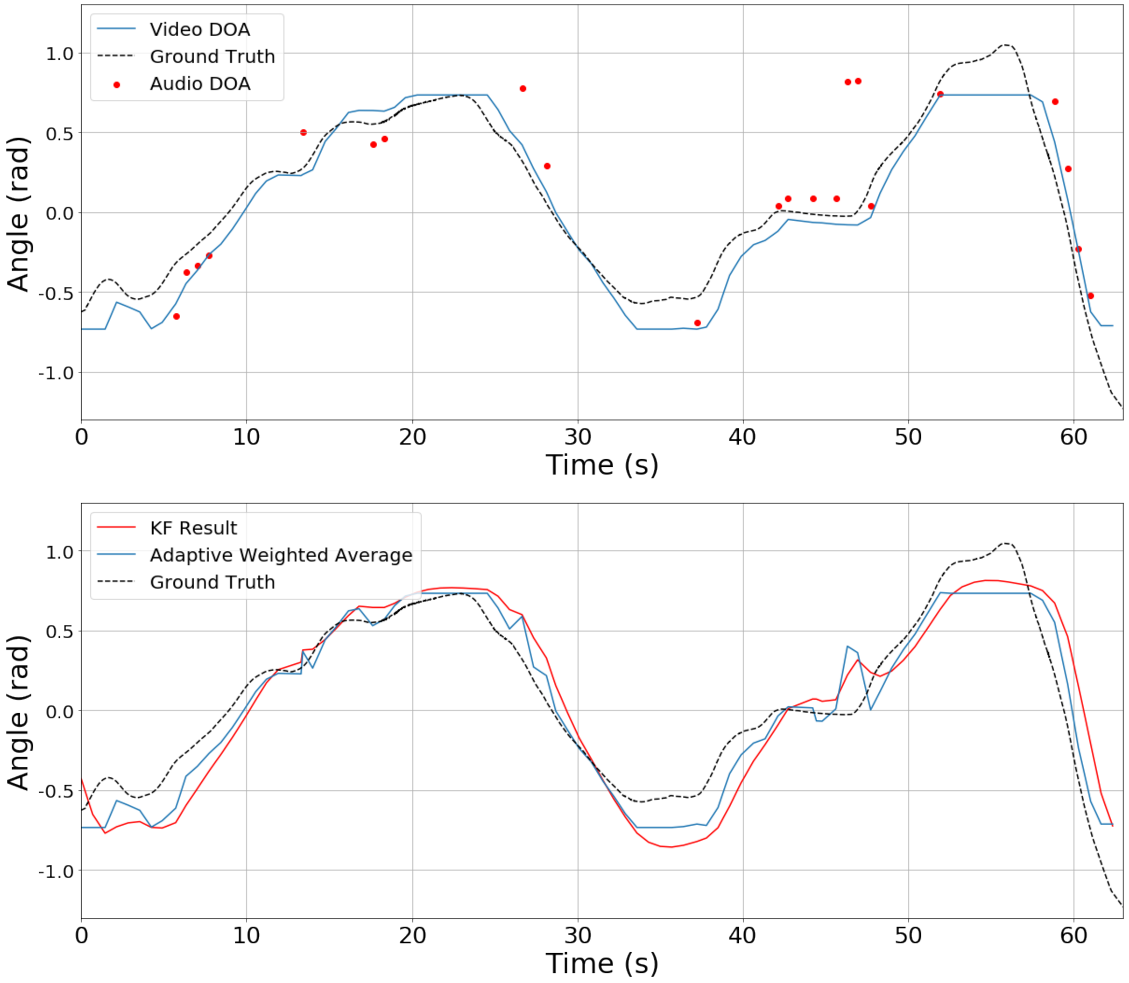

Figure 5.9: Video and audio estimates with ground truth (top) shown alongside adaptive weighted average and Kalman filtered responses (below)

the sensor fused audio and video approach in the scene of following a single moving human subject. Figure 5.1 provides an overview of the process behind each approach. Figure 5.11 shows the behaviour and response of the REEM-C with a moving conversational partner in a similar trial. Figure 5.10 shows average scores for the criteria with both approaches, along with standard error placed as error bars.

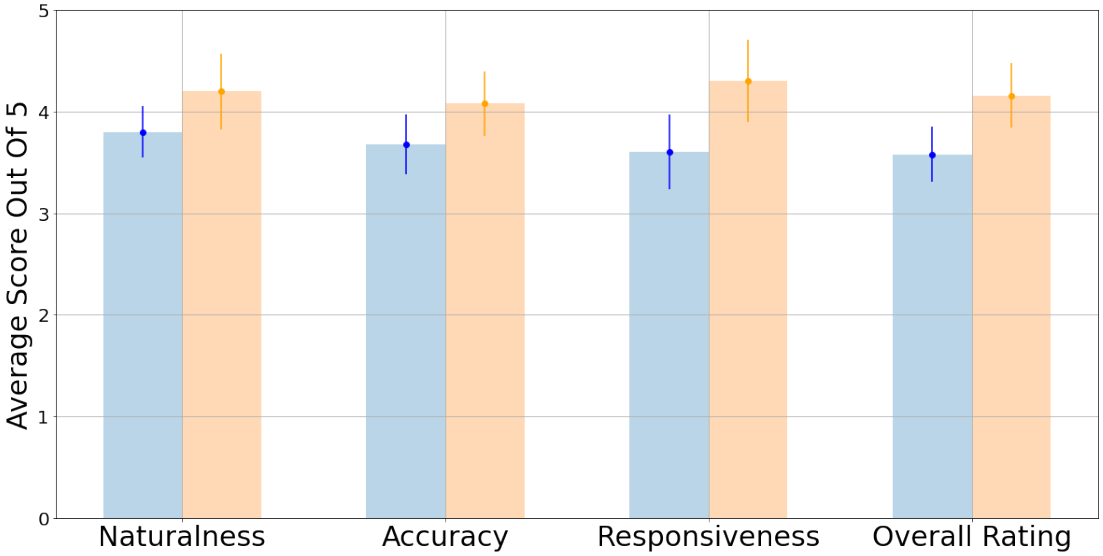

Figure 5.10: Visualization of average criterion scores for both approaches. Blue = audio only, orange = fused approach

The aim of this small study was to investigate the initial perception of using both approaches. These preliminary results, although not statistically significant, suggest that the average scores were higher for the condition where sensor fusion was used. Since the audio DOA exceeds the field of view of the video, the fused method is also more robust, whether or not the participants are able to immediately notice the difference.

We aim to extend our current work by evaluating the multi-subject scenario in greater detail with all proposed systems utilized in tandem, which will test the behaviour for alternating the robot's attention between subjects who are both inside and outside the field of view. This extension may require further capabilities to identify when subjects are speaking to one another as opposed to the robot. This may be achieved via estimation of the subjects' head pose, or by using gaze estimation as described in [3]. Figure 5.12 shows a visual representation of the robot's position, field of view, and the visual and audio information it perceives of a moving human subject.

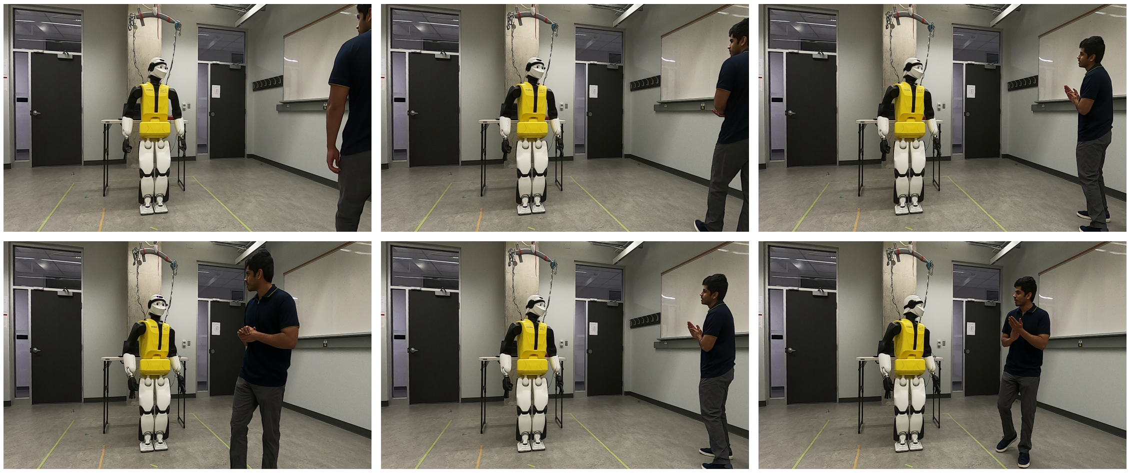

Figure 5.11: Frames depicting the REEM-C moving its head, shoulders and feet to orient to a moving conversational partner (view from top left to bottom right)

5.8 Discussion

Overall, the audio and video intelligence presented work robustly. The loud fans and motor noises of the REEM-C do not interfere with the performance of the audio DOA estimation. Video DOA estimation is also reliable and agrees significantly with results from audio. The sensor fusion approach along with a Kalman filter intelligently combines received information and provides natural moving targets for the REEM-C to follow in real-time.

As shown in Table 5.2, the accuracy for all trials are not similar. Trials 4, 5 and 6 involved two participants having dynamic conversations and talking over each other, explaining the drop in accuracy compared to the first 3 trials. However, Table 5.3 indicates a fairly high precision for those trials, indicating the chance for false positives is low even in dynamic conversational scenarios. Overall, this is preferable as it ensures the robot is less likely to reorient itself in an unpredictable manner or at incorrect times.

The user study results consistently show that tracking with the sensor fused approach is overall reliable. Using both sensors also facilitates tracking between two or more potential speakers. These preliminary results from the subjects indicate that using the proposed approach generates natural and responsive motions, providing some initial insights into how humans may perceive such a system if it were to be deployed in an HRI context.

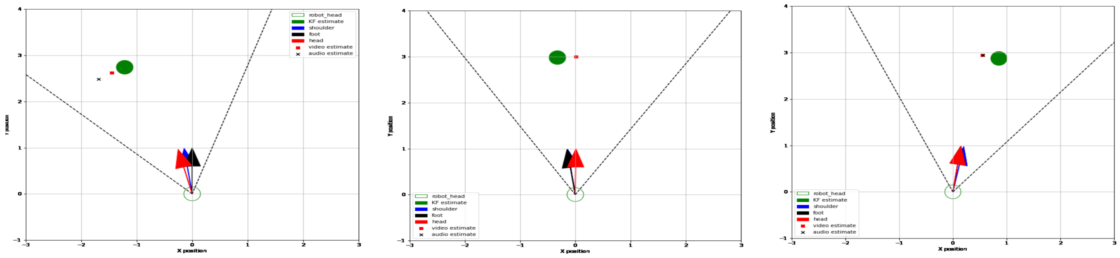

Figure 5.12: Bird's eye animation of REEM-C, along with its field of view shown in black dotted lines, depicting a visually detected conversational partner shown as the circle. The circle is filled with a red dot when the audio DOA is closely associated with the visual DOA of the subject

5.9 Conclusions

In this paper, we demonstrate the development and integration of audio and video intelligence algorithms for the purposes of more complete human-robot interaction in difficult scenarios. Our algorithms are shown to work accurately on the robot and generate good subject tracking results. Qualitative results from users show that the HRI experience was realistic and accurate, opening the door for more encompassing HRI frameworks. Future work may aim to evaluate the behaviour of the robot with multiple subjects and utilize the extended visual intelligence generated by the presented algorithms to develop more robust HRI platforms.

Chapter 6

Chatbot System

This section briefly outlines the work done on integrating a conversational pipeline for the REEM-C in order to complete the human-robot interaction loop. The pipeline involves a real-time speech-to-text functionality, response generation using a large-language model (LLM) and a text-to-speech feature to communicate again with the interlocutor. This system tentatively completes the human-robot interaction framework as it allows for the REEM-C to not only track moving human partners, but also communicate with them to a relatively intelligent degree.

6.1 System Components

The systems outlined in previous chapters lay the framework for the developed chatbot system. The audio stream used for DOA estimation is sent to a separate ROS node to perform transcription. A flowchart in Figure 6.1 briefly outlines the chatbot pipeline and how it connects to the previously defined systems. The same audio stream used for the detection and tracking system is fed into the chatbot pipeline, allowing for a seamless human-robot interaction framework.

The buffer of audio data for speech recognition is of size 20, and given the collected frame size of 350ms from Chapter 2, results in a potential 7 seconds of audio that can be transcribed at any given time. Transcriptions are performed via a Google speech recognition service, easily accessed via Python API. In order to maintain conversational fluidity and minimize time spent collecting data or transcribing, a simple frame-to-frame energy detector is used to identify when speech has begun and speech has ended. During silence,

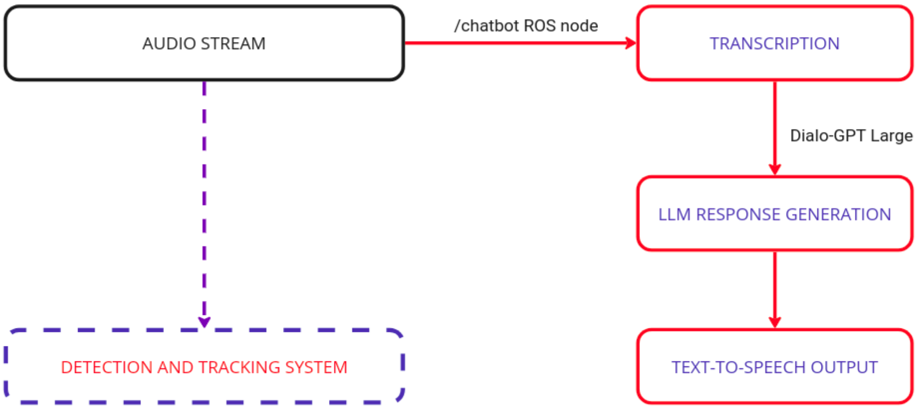

Figure 6.1: Usage of chatbot system given recorded audio stream

if a significant amount of energy is present in the next frame (in terms of a ratio of frame n+1 to n), human speech is said to have begun. Once the energy ratio of the frame n and n-1 is low, the speech is marked as complete. From the audio buffer, only the frames where the speech is considered present is sent to the transcription service, to optimize efficiency of the pipeline.

The transcription is generated in the separate ROS node, such that the other systems can continue to run in parallel, allowing the REEM-C to continue tracking detected interlocutors. The transcription is then passed into an LLM designed to generate conversational responses. The LLM is a pre-trained conversational transformer from HuggingFace, called the Dialo-GPT Large. Free trial of Chat-GPT by OpenAI also proved to be a worthwhile choice for conversational interactions. The choice of LLM at this stage is arbitrary and determined based on the quality of responses generated on a few sample inputs. This LLM generates responses to given text inputs from the transcription step, facilitating a conversational human-robot interaction scenario.

Finally, the generated response is sent back via text-to-speech. This allows for the REEM-C and any human subject to converse with each other in a regular manner, while the REEM-C also tracks the human in the ways described in previous chapters.

6.2 Implementation Details and Latency

A short evaluation is performed for the individual components and how much time they take individually, helping characterize the overall time taken for the system to respond once an input is given. A series of speech inputs are recorded, and the time taken for each step is presented in Table 6.1 below. The total response time column is the time it takes for the system to begin responding with the text-to-speech, and is the sum of the time it takes to transcribe and generate the response. This response time thus represents how long an interlocutor would have to wait to get a response once their speech has been received.

Table 6.1: Time Taken For Chatbot System Given Speech Input

Input Text	**Transcription (s)**	**Response Generation (s)**	**Total Response (s)**
hello there	0.83	0.44	1.27
what is your name	1.19	1.07	2.26
can you answer that again	1.22	0.62	1.84
how are you	1.25	0.85	2.10
now I am trying to finish this conversation	1.18	1.07	2.25

It can be seen that each system is not fast enough to generate a realistic response time that conforms to human conversational behaviour. Regardless of the length of the input speech, the time taken to transcribe hovers around 1 second on average. Generating and delivering the response also takes a similar amount of time, resulting in a system that takes much longer than necessary.

In order to smoothen this process, we introduce the use of filler words. The goal is to use filler words such as ”hmm”, ”I see”, and ”ahh”, that can be said while the above pipeline runs, and that reduces the perceived latency of the conversational interaction. Filler words are shown to increase the social presence of robots [12], and therefore have a direct value in this presented scenario. A diagram depicting this is shown in Figure 6.2 below.

Once speech is determined to be complete, a filler word is inserted after about 200ms, so the naturalness of the conversation can be maintained. This occurs while the chatbot

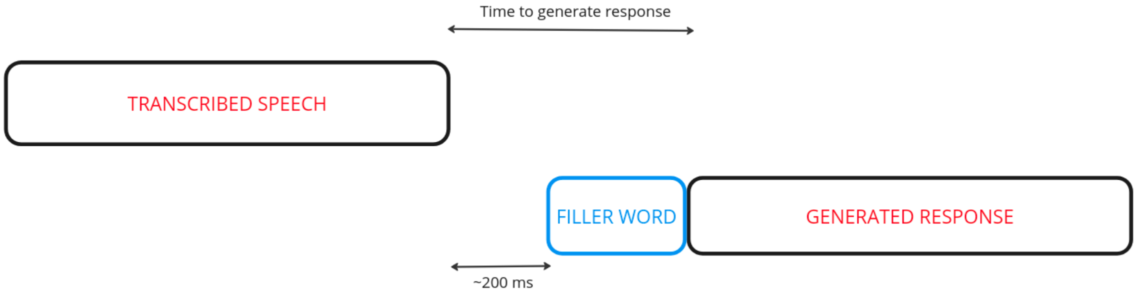

Figure 6.2: Placement of filler words in conversational pipeline

system is transcribing and generating a response, and so the perceived latency for the conversation can be significantly reduced. The use of gestural fillers has been shown to reduce this perceived latency [22], and so the vocal fillers may be well accompanied by corresponding gestural fillers such as leaning, nodding, or more complex postures with the body and arms.

6.3 Summary

Overall, this chatbot system lays out the groundwork for how the REEM-C can engage in a complete verbal interaction while also following its interlocutors. We have presented a system that integrates existing tools for speech recognition and conversational response generation, while evaluating their timing requirements and presenting ways in which they can be ameliorated. Future work may consider designing gestures with the upper body and arms that correspond to the content of the conversation, to add another layer of realism. Faster methods for speech transcription and good conversational response generation should also be explored.

Chapter 7

Conclusion

7.1 Summary

In this thesis, a variety of robotic systems were developed to improve human-robot interaction. These include the acoustic sound source localization algorithms, the complementary computer vision algorithms, the humanoid tracking model, and the integration of the above into one comprehensive HRI framework. Algorithmic performance was evaluated on collected quantitative data in terms of accuracy and latency. Successful real time deployment on the REEM-C of the above systems validates the proposed methods and presented results.

Audio DOA Pipeline

A pipeline was developed to perform real-time binaural direction of arrival estimation. This pipeline involved a method to identify human speech with spectral analysis, and measure timing differences to accurately generate an angle of arrival of sound. To optimize performance, a dataset was collected and annotated. The dataset was used to choose the best parameters for the pipeline using a brute force and Bayesian parameter optimization method, which ensured high accuracy and low latency for real-time estimation.

Vision Systems for HRI

A series of lightweight computer vision algorithms were also developed to estimate the behaviour of detected interlocutors in the REEM-C's field of view. These algorithms were developed using a facial landmark detector and pattern recognition techniques. These include visual voice activity detection, gaze direction estimation, and visual direction estimation (which complements the acoustic DOA pipeline). Algorithms are determined to work fast and accurately

Humanoid Tracking Behaviour

A motion capture study was conducted, which provided insight into how humans may track moving conversational partners. The measured participant data was modeled by inspecting the participant's head, shoulders and feet behaviour. A mathematical model was developed to model this behaviour, and used on the REEM-C to facilitate realistic, humanlike tracking behaviour.

Sensor Fusion and System Integration

Additionally, the developed systems were integrated for real time usage on the REEM-C. Audio and visual DOA were combined using a sensor fusion approach, allowing participants to be smoothly tracked with the developed tracking model. A feature to facilitate reorientation was also developed, allowing for the REEM-C to maintain interaction with multiple human subjects at once. A user study evaluates and confirms that interaction performance is improved when the sensor fused and integrated system is used.

Chatbot System

Finally, initial work to integrate a chatbot system for the REEM-C was carried out using speech recognition, a conversational LLM and a text-to-speech feature. The drawbacks of such an integration were explored, and future steps for improving and completing the conversational loop for human-robot conversation were proposed.

7.2 Future Work

Future work may involve augmenting the behaviours of the humanoid robot for more dynamic human robot interaction. This may include gestures with the arms and hands, and more expressive movements with the body. Extended work could be done to model this behaviour with more motion capture analysis, which may add another layer of intelligent human-robot interaction.

The acoustic DOA pipeline may be improved by ameliorating the front-back ambiguity. This is difficult to do binaurally, so a third microphone as part of the array may be useful. This can allow the REEM-C to turn its attention to humans standing behind it as well, which further opens up its capabilities and possibilities for HRI. A conversational framework may also be useful for HRI using both DOA estimation and speech recognition technology. The proposed chatbot system can then allow the robot to properly communicate with humans, given further work to optimize the time it takes to transcribe and generate responses.

Visual systems can ideally be replaced by more powerful and accurate techniques. These works may involve machine learning tools to perform similar tasks, or extensions, such as pose estimation and gesture detection. These techniques may further augment the acoustic intelligence and allow the robot to identify and respond to more complex HRI scenarios.

References

[1] Alexander Bannat, Jürgen Gast, Tobias Rehrl, Wolfgang Rösel, Gerhard Rigoll, and Frank Wallhoff. A multimodal human-robot-interaction scenario: Working together with an industrial robot. pages 303–311, 01 1970.

[2] Pranav Barot, Ewen MacDonald, and Katja Mombaur. Natural head and body orientation for humanoid robots during conversations with moving human partners through motion capture analysis. 2023.

[3] Pranav Barot, Ewen MacDonald, and Katja Mombaur. Vision systems for identifying interlocutor behaviour and augmenting human robot interaction. In *Conference On Computational Vision And Imaging Systems*, volume 8, page 55–58, May 2023.

[4] Toufik Bentaleb and L. Shahin. Humanoid robots imitation of human motion using off-line and real-time adaptation techniques. *Int. J. Mechanisms and Robotic Systems*, Vol. 2, Nos. 3/4, 2015:295–313, 01 2015.

[5] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science Discovery*, 8(1):014008, 2015.

[6] Karsten Berns and Zuhair Zafar. Emotion based human-robot interaction. *MATEC Web of Conferences*, 161:01001, 01 2018.

[7] Angelo Cangelosi and Tetsuya Ogata. Speech and language in humanoid robots. In *Humanoid Robotics: A Reference.*

[8] Aaron Chau, Kouhei Sekiguchi, Aditya Arie Nugraha, Kazuyoshi Yoshii, and Kotaro Funakoshi. Audio-visual slam towards human tracking and human-robot interaction in indoor environments. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–8, 2019.

[9] Ruth E. Corps, Birgit Knudsen, and Antje S. Meyer. Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation. *Cognition*, 223:105037, 2022.

[10] B. Dariush, M. Gienger, B. Jian, C. Goerick, and K. Fujimura. Whole body humanoid control from human motion descriptors. 2008.

[11] D. Fraser and J. Potter. The optimum linear smoother as a combination of two optimum linear filters. *IEEE Transactions on Automatic Control*, 14(4):387–390, 1969.

[12] Henry Goble and Chad Edwards. A robot that communicates with vocal fillers has . . . uhhh . . . greater social presence. *Communication Research Reports*, 03 2018.

[13] Sylvain Guy, Stéphane Lathuilière, Pablo Mesejo, and Radu Horaud. Learning visual voice activity detection with an automatically annotated dataset. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4851–4856. IEEE, 2021.

[14] Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

[15] Marina Horn, Manish Sreenivasa, and Katja Mombaur. Optimization model of the predictive head orientation for humanoid robots. 2015:767–772, 02 2015.

[16] Takao Imai, Steven Moore, Theodore Raphan, and Bernard Cohen. Interaction of the body, head, and eyes during walking and turning. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 136:1–18, 02 2001.

[17] Choi K.-J. and Ko H.-S. Online motion retargetting. In *The Journal of Visualization and Computer Animation*, 2000.

[18] Davis King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 07 2009.

[19] Marius Klug and Andreas Zell. Emotion-based human-robot-interaction. 07 2013.

[20] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[21] Kenji Koide, Emanuele Menegatti, Marco Carraro, Matteo Munaro, and Jun Miura. People tracking and re-identification by face recognition for rgb-d camera networks. pages 1–7, 09 2017.

[22] Junyeong Kum and Myungho Lee. Can gestural filler reduce user-perceived latency in conversation with digital humans? *Applied Sciences*, 12(21), 2022.

[23] Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 2015.

[24] Riccardo Levorato and Enrico Pagello. Doa acoustic source localization in mobile robot sensor networks. In *2015 IEEE International Conference on Autonomous Robot Systems and Competitions*, pages 71–76, 2015.

[25] Katrin Lohan, Hagen Lehmann, Christian Dondrup, Frank Broz, and Hatice Kose. *Enriching the Human-Robot Interaction Loop with Natural, Semantic, and Symbolic Gestures*, pages 1–21. 09 2017.

[26] Katrin Lohan, Hagen Lehmann, Christian Dondrup, Frank Broz, and Hatice Kose. *Enriching the Human-Robot Interaction Loop with Natural, Semantic, and Symbolic Gestures*, pages 1–21. 01 2017.

[27] Quang Nguyen, Sang-Seok Yun, and Jongsuk Choi. Audio-visual integration for human-robot interaction in multi-person scenarios. 09 2014.

[28] Shokoofeh Pourmehr, Jack Thomas, Jake Bruce, Jens Wawerla, and Richard Vaughan. Robust sensor fusion for finding hri partners in a crowd. pages 3272–3278, 05 2017.

[29] Ewen MacDonald Pranav Barot, Katja Mombaur. Estimating speaker direction on a humanoid robot with binaural acoustic signals. *Journal Of Public Library Of Science (PLOS ONE)*, 2023.

[30] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210, 08 2017.

[31] Caleb Rascon, Ivan Meza, Gibran Fuentes-Pineda, Lisset Salinas, and Luis Pineda. Integration of the multi-doa estimation functionality to human-robot interaction. *International Journal of Advanced Robotic Systems*, 12, 02 2015.

[32] Neelesh Rastogi, Fazel Keshtkar, and Md Suruz Miah. A multi-modal human robot interaction framework based on cognitive behavior therapy model. 07 2018.

[33] M. Risoud, J.-N. Hanson, F. Gauvrit, C. Renard, P.-E. Lemesre, N.-X. Bonne, and C. Vincent. Sound source localization. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 135(4):259–264, 2018.

[34] Anwar Saeed, Ayoub Al-Hamadi, and Michael Heuer. Multi-modal fusion framework with particle filter for speaker tracking. *International journal of future generation communication and networking. - Taejŏn : SERSC, Bd. 5*, 01 2012.

[35] Anwar Saeed, Ayoub Al-Hamadi, and Michael Heuer. Speaker tracking using multi-modal fusion framework. In Abderrahim Elmoataz, Driss Mammass, Olivier Lezoray, Fathallah Nouboud, and Driss Aboutajdine, editors, *Image and Signal Processing*, pages 539–546, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[36] Giulio Sandini, Alessandra Sciutti, and Francesco Rea. *Movement-Based Communication for Humanoid-Human Interaction*, pages 2169–2197. 01 2019.

[37] Giulio Sandini, Alessandra Sciutti, and Francesco Rea. *Movement-Based Communication for Humanoid-Human Interaction*, pages 2169–2197. 01 2019.

[38] João F. Santos, Mohammed Senoussaoui, and Tiago H. Falk. An improved non-intrusive intelligibility metric for noisy and reverberant speech. In *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 55–59, 2014.

[39] Marco Sewtz, Tim Bodenmüller, and Rudolph Triebel. Sound source localization for robotic application. 06 2020.

[40] Zhanbo Shi, Lin Zhang, and Dongqing Wang. Audiondash;visual sound source localization and tracking based on mobile robot for the cocktail party problem. *Applied Sciences*, 13(10), 2023.

[41] Vanlanduit S. Shi L, Copot C. Gaze gesture recognition by graph convolutional networks. In *Front Robot AI. 2021 Aug 5*, 2021.

[42] Spyridon Siatras, Nikos Nikolaidis, Michail Krinidis, and Ioannis Pitas. Visual lip activity detection and speaker detection using mouth region intensities. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(1):133–137, 2009.

[43] Ludwig Sidenmark and Hans Gellersen. Eye, head and torso coordination during gaze shifts in virtual reality. volume 27, pages 1–40. ACM New York, NY, USA, 2019.

[44] M.N. Sreenivasa, Philippe Souères, and Jean-Paul Laumond. Walking to grasp: Modeling of human movements as invariants and an application to humanoid robotics. *IEEE Transactions on Systems, Man, and Cybernetics - TSMC*, 42:880–893, 07 2012.

[45] Tanya Stivers, N. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan De Ruiter, Kyung-Eun Yoon, and Stephen Levinson. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 106:10587–92, 07 2009.

[46] Raquel Viciana-Abad, Rebeca Marfil, Jose Perez-Lorenzo, Juan Bandera, Adrián Romero-Garcés, and P. Reche-López. Audio-visual perception system for a humanoid robotic head. *Sensors*, 14:9522–9545, 06 2014.

[47] Caiyong Wang, Yunlong Wang, Yunfan Liu, Zhaofeng He, Ran He, and Zhenan Sun. Sclerasegnet: An attention assisted u-net model for accurate sclera segmentation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(1):40–54, 2020.