Luís Miguel da Silva Costa

Integration of a Communication System for Social Behavior Analysis in the SocialRobot Project

October/2013

Universidade de Coimbra

University of Coimbra

Faculty of Science and Technology

Department of Electrical and Computer Engineering

Integration of a Communication System for Social Behavior Analysis in the SocialRobot Project

Luís Miguel da Silva Costa

A dissertation presented for the degree of Master of Science in Electrical and Computer Engineering

October, 2013

University of Coimbra

Faculty of Science and Technology

Department of Electrical and Computer Engineering

Integration of a Communication System for Social Behavior Analysis in the SocialRobot Project

Supervisor

Dr. Jorge Manuel Miranda Dias

Co-Supervisor

Eng. Pedro Emanuel dos Santos Vale Sousa Trindade

Juries

Dr. Rui Alexandre de Matos Araújo

Dr. Luís Alberto da Silva Cruz

Dissertation presented to the Electrical and Computer Engineering Department of the Faculty of Science and Technology of the University of Coimbra in partial fulfillment of the requirements for the Degree of Master of Science in Electrical and Computer Engineering

October, 2013

Acknowledgments

I want to thank my supervisor, Dr. Jorge Manuel Miranda Dias, for the opportunity to do this thesis and his guidance, which always pointed me into the right direction.

I also acknowledge my co-supervisor, Eng. Pedro Trindade, for his patience, advices and neverending availability to help me during this dissertation. His guidance caused me to evolve my critical spirit, turning me into a much more capable scientist.

I can't forget to thank all my colleagues at the Mobile Robotics Laboratory (MRL) from the University of Coimbra's Institute of Systems and Robotics (ISR) for their fellowship and motivation.

At last but not least, i am grateful to my family and friends for their neverending support and for always believing in me, even in times of doubt.

Luís Miguel da Silva Costa, October 2013

Abstract

Human-Robot-Interaction is a vast research field with great potential and importance for the assignment of communication capabilities to robots.

Dialogs can be considered as one of the most common ways of communication between humans. Through the auditory system, humans can locate the person with whom they are dialoguing, listen to what is said and perceive the emotional state of that person. In order to deliver robots with those same capabilities, it is necessary to implement an artificial auditory system, with functionalities similar to the human's auditory system. In this thesis a integration effort is made with a module for the sound source localization, a speech recognition module and a module for the voice emotion recognition in order to achieve that goal.

This work is based on capturing sound through microphones. For this purpose we use a sound processing toolkit named HARK, which enables the estimation of the location of a sound source using the MUltiple SIgnal Classification (MUSIC) method; a speech recognition toolkit, the CMU-Sphinx, using its decoder Pocketsphinx, which uses a Hidden Markov Model classifier (HMM) and finally a toolkit for the voice emotion recognition named openEar which uses a Support Vector Machine classifier (SVM). This thesis also presents a module for the visualization of the sound source localization results and synthesis modules, such as a speech synthesizer and an application launcher.

All modules in this thesis are integrated under the Robot Operating System (ROS), which gives the system a modular feature, providing a relative independence between them.

In order to validate the communication system in this work, we considered some use cases scenarios, which are based on the SocialRobot project, an european project which has the University of Coimbra as a development partner.

The integration also delivers an action controller module to manage the steps of the use case scenarios, which performs the corresponding actions according to the user's social behavior. The social behavior is determined by combinations of the speech and emotion recognition results, which are associated to pre-defined actions for each use case scenario.

Key Words: Human-Robot-Interaction (HRI); Sound Source Localization (SSL); Speech Recognition; Voice Emotion Recognition; Robot Operating System (ROS); SocialRobot Project.

Resumo

A interacção entre robots e humanos (HRI) é um vasto campo de investigação com enorme potencial e importância para a atribuição de capacidades comunicativas aos robots.

Uma das mais usuais vias de comunicação entre pessoas é o dialogo. Através do seu sistema auditivo, os humanos conseguem localizar a pessoa com a qual estão a dialogar, ouvir aquilo que é dito e perceber qual a emoção transmitida por essa pessoa. De forma a conseguir que os robots tenham essas mesmas capacidades, é necessária a implementação de um sistema auditivo artificial, semelhante ao humano nas suas funcionalidades.

Nesta tese é feita a integração de um módulo de localização da fonte de som, um módulo de reconhecimento do discurso e um módulo de reconhecimento da emoção, de modo a atingir esse objectivo.

Este trabalho baseia-se na captura de som através de microfones. Para esse efeito é utilizada uma toolkit para o processamento de som denominada por HARK, a qual estima a localização de uma fonte de som através do método *MUltiple SIgnal Classification* (MUSIC); uma toolkit para o reconhecimento do discurso denominada por *CMU-Sphinx*, fazendo uso de uma das suas ferramentas de reconhecimento, o *Pocketsphinx*, que efectua o reconhecimento de palavras através de um classificador por Hidden Markov Models (HMM); por fim é usada uma toolkit para o reconhecimento da emoção presente na voz denominada por openEar que efectua a reconhecimento usando um classificador por *Support Vector Machine* (SVM). Nesta tese é ainda apresentado um módulo para a visualização dos resultados da localização da fonte de som e módulos de síntese, para simulação de voz e execução de aplicações.

Todos os módulos apresentados nesta tese são integrados dentro do *Robot Operating System* (ROS), o que atribui ao sistema uma característica modular, proporcionando uma independência relativa entre os módulos.

De forma a validar sistema de comunicação deste trabalho são considerados alguns casos de uso, os quais têm como base os cenários do projecto europeu SocialRobot, que contempla a Universidade de Coimbra como um dos seus parceiros.

Da integração resulta ainda um módulo para o controlo da interação descrita nos casos de uso, que identifica e inicia acções de acordo com o comportamento social dos utilizadores. O comportamento do utilizador é determinado atavés da combinação dos resultados do reconhecimento da fala e da emoção, associados a ações predefinidas para os casos de uso.

Palavras chave: Interacção entre pessoas e robots (HRI); Localização da fonte de som (SSL); Reconhecimento da fala; Reconhecimento da emoção; Robot Operating System (ROS); SocialRobot Project

Declaration

The work in this thesis project is based on research carried out at the Mobile Robots Laboratory of the Institute of Systems and Robotics - University of Coimbra. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Luís Miguel da Silva Costa, October 2013

Contents

Acknowledgements i

Abstract ii

Resumo iii

Declaration iv

List of Acronyms ix

List of Figures xi

List of Tables xii

1 Introduction 1

1.1 Context and Motivation 1

1.2 Human-Robot-Interaction 2

1.3 SocialRobot Project 2

1.4 Approach 3

1.5 Outline of the document 5

2 State of the Art 7

2.1 Sound Source Localization 7

2.1.1 Introduction 7

2.1.2 Related Work 8

2.1.3 Toolkits 11

2.2 Speech Recognition 12

2.2.1 Introduction 13

2.2.2 Related work 13

2.2.3 Toolkits 16

2.3 Voice Emotion Recognition 17

2.3.1 Introduction 17

2.3.2 Related Work 18

2.3.3 Toolkits 20

3 Use Case Scenarios 21

3.1 Introduction 21

3.2 Use Case Scenario 1 21

3.3 Use Case Scenario 2 23

3.4 Use Case Scenario 3 24

3.5 Discussion 26

4 Implementation and Integration 29

4.1 Framework 29

4.2 Robot Operating System 30

4.2.1 Introduction 30

4.2.2 ROS Integration Framework 31

4.3 Sound Capture Package 33

4.4 Sound Source Localization Package 34

4.4.1 HARK 35

4.5 Speech Recognition Packages 36

4.5.1 CMU-Sphinx 37

4.6 Voice Emotion Package 39

4.6.1 OpenEar 39

4.7 Synthesis Packages 40

4.8 System Integration 41

4.8.1 Voice Command Selection Package 42

4.8.2 Emotion Selection Package 43

4.8.3 Action Controller Package 43

5 System Results **45**

5.1 Individual Results 45

5.2 Integration Results 49

6 Discussion and Future Work **53**

6.1 Discussion 53

6.2 Future Work 54

6.2.1 Improvements 55

6.2.2 Beyond Auditory Analysis 56

7 Conclusions **57**

Bibliography **59**

List of Acronyms

AANN	Auto Associative Neural Networks
ABC	Airplane Behavior Corpus
AFE	Advanced Front-End
AVIC	Audio Visual Interest Corpus
BVM	Bayesian Volumetric Map
CASA	Computational Auditory Scene Analysis
CM	Correlation Matrix
CMVN	Cepstral Mean and Variance Normalization
CSP	Cross-power Spectrum Phase
DOA	Direction of Arrival
DSBF	Delay and Sum Beam Forming
EMO-DB	Berlin Speech Emotion Database
FFT	Fast Fourier Transform
FP7	Seventh Framework Programme
GEVD	Generalized Eigen Value Decomposition
GMM	Gaussian Mixture Model
GPL	General Public License
GSVD	Generalized Singular Value Decomposition
GUI	Graphical User Interface
HEQ	Histogram EQualization
HMM	Hidden Markov Model
HOCMN	Higher Order Cepstral Moment Normalization
HRI	Human-Robot-Interaction
IAPP	Industry-Academia Partnerships and Pathways
ILD	Inter-aural Level Differences
ITD	Inter-aural Time Differences
k-NN	k-Nearest Neighborhood
LDC	Linear Discriminant Classifier
LLD	Low-Level Descriptors
MFCCs	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood

MUSIC	MUltiple SIgnal Classification
MVDR	Minimum Variance Distortionless Response
PCA	Principal Component Analysis
PMC	Parallel Models Combination
ROS	Robot Operating System
SAL	Sensitive Artificial Listener
SEVD	Standard Eigen Value Decomposition
SGMM	Subspace Gaussian Mixture Models
SLAM	Speaker Localization by an Arrayed Microphone
SMILE	Speech and Music Interpretation by Large-space Extraction
SNR	Signal-to-Noise Ratio
SRP-PHAT	Steered Response Power using the PHAse Transform
SSL	Sound Source Localization
SVM	Support Vector Machine
VAM	Vera am Mittag
VC	Voice Characterization
VCRR	Voice Command Recognition Rate
WCC	Weighted Cross Correlation
WSJ	Wall Street Journal

List of Figures

1.1 Module organization for integration and social behavior analysis. 4

2.1 Sound Source Localization - Sound's direction of arrival. 8

2.2 HARK sound source localization for four simultaneous talkers in a 360º range. 12

2.3 Recognition of a speech utterance. 14

2.4 Pocketsphinx ROS package recognizing the “hello” word from a live speech signal. 17

2.5 Emotion recognition from audio input. 18

2.6 openEAR live emotion recognition from speech signals using the pre-trained models and LibSVM library. 20

3.1 Elderly woman lying on the floor 22

3.2 Flowchart of the Use Case Scenario 1 22

3.3 Elderly in bed, assisted by the centre staff called by the SocialRobot 23

3.4 Flowchart of the Use Case Scenario 2 24

3.5 Left side - Elderly socializing on the centre living room. Right side - Medication brought by the Social Robot. 25

3.6 Flowchart of the Use Case Scenario 3 26

4.1 Framework diagram 29

4.2 Analysis Diagram 30

4.3 Synthesis Diagram 30

4.4 ROS nodes communication through a ROS topic. 31

4.5 ROS integration diagram. 32

4.6 ROS message types and contents. 33

4.7 Sound Capture Network 33

4.8 Sound Source Localization Network 34

4.9 Graphical representation of a transfer function file, containing 72 possible sound sources for the SSL task, created with the harktool4. 35

4.10 Sound Source Localization results display. 41

4.11 Left side - Best angle range for each Kinect's microphone. Right side - Kinect microphone array disposal. 42

4.12 Graphical representation of the affective circumflex, with the valence dimension represented horizontally on the x-axis and arousal vertically on the y-axis [3]. 43

5.1 Sound Source Localization results. 46

5.2 Speech recognition overall recognition rates. 47

5.3 Simulation of the use case scenario #1. 50

5.4 Simulation of the use case scenario #2. 51

5.5 Simulation of the use case scenario #3. 52

6.1 Visual system framework 56

List of Tables

4.1 Parameter list for the LocalizeMUSIC module 36

4.2 Low-Level Descriptors in openEAR's SMILE feature extractor. 40

4.3 Use Case Scenario 1 Interaction Stages 44

4.4 Use Case Scenario 2 Interaction Stages 44

4.5 Use Case Scenario 3 Interaction Stages 44

5.1 Results for Sound Source Localization. 46

5.2 Voice command recognition for speech signals from specific angles. VCRR (Voice Command Recognition Rate) 47

5.3 Confusion table of the emotion recognition results for the EMO-DB model, including seven emotion classes. 48

5.4 Confusion table of the emotion recognition results for the arousal-valence approach. 49

5.5 Success rates for the use case scenarios simulations. 50

Chapter 1

Introduction

1.1 Context and Motivation

Social robots popularity has increased along the years, which lead them to be considered to integrate human's social lives. Social robots are now an emergent research field, with the ultimate goal of providing solutions and designs, suitable for human's needs. Many roles for social robots have been discussed all over the years, but more recently some projects and developments have been made in order to create service social robots. Service social robots can be integrated in various scenarios, but with the increasing numbers of the elderly population, elderly service and assistance has become a main target of intervention.

Many studies and research have been done in order to infer the social acceptance by human's society. For that acceptance, many issues have been a case study, such as the robot physiognomy, communication skills and cultural adaptation's. Those studies indicate how likely a person will accept a robot's interaction, depending on its characteristics, his culture and his expectations about the interaction itself. Moreover, if the interaction with a robot presents a challenge, people tend to be discouraged to interact with it.

Usually humans communicate trough dialogs, implying the ability to speak and hear. Thus robots need to have similar capabilities, and the most usual way to provide them is trough speech analysis and synthesis modules.

The design and development of a social robot should deliver communication and interaction skills towards an effortless communication, similar to the human's communication patterns.

1.2 Human-Robot-Interaction

Human-Robot-Interaction (HRI) is a field of study dedicated to understanding, designing, and evaluating robotic systems for use by or with humans. Interaction, by definition, requires communication between robots and humans[16]. Interaction with robots can be achieved in many ways, depending on the type of robot we intend to interact with. Since the beginning of robotics research, many types of robots have been designed. For industrial purposes we have robot manipulators, most commonly known as robot arms, which are intended to perform a repetitive task taught by humans. With the evolution of that type of robot, in terms of precision and accuracy, it was consider to be used for medical purposes, such as surgery teleoperated manipulators. More recently robot designs have a tendency for being biologically inspired, and the most representative category for such design is the humanoid robots. The intentions for the use of robots have also changed along the years. Nowadays robots are starting to be accepted by human's society and they have passed on to being considered as possible collaborators, assistants and companions, for human's daily tasks and needs, instead of being simply a machine that only serves one purpose, such as performing a simple task. With this new possible role for robots in human's lives, arrangements have to be made in order to fulfill human's expectations. The most important of those expectations can be described as social skills for interaction. For a simple and reliable interaction between humans and robots, the most obvious requirements are communication and knowledge. In terms of communication, many approaches have been developed over the years, but the most accepted are those considering a naturalistic way of communication similar to the human's social behavior. To accomplish that kind of communication, robots need to have knowledge that gives them the means to understand humans. That knowledge can be referred to the social context and surroundings, the informations passed trough the communication process or even the knowledge of the human's social behavior. To acquire those informations, robots are equipped with sensors, such as microphones and cameras that simulate human's auditory and vision senses. Just like human's have central processing unit that we call brain, robots are also required to have one in order to analyze the information captured by their sensors. The ability to analyze sensorial information and react accordingly to a social behavior profile, gives a robot what is commonly known as artificial intelligence.

1.3 SocialRobot Project

According to demographic studies, the Europe's elderly population will double by middle of this century, bringing us the necessity create solutions and develop strategies to deal with this demographic change [33]. As people grow older, their physical and cognitive capabilities tend to decrease, which might lead them to a situation where their are not fully autonomous. Thus, in order to maintain elderly's quality of life, they need to stay independent and connected with their social environment and society in general for as long as possible.

SocialRobot [7] is an IAPP (Industry-Academia Partnerships and Pathways) Marie Curie Actions european project, within the FP7 (Seventh Framework Programme), coordinated by the University of Coimbra and includes a consortium formed by four partners from two EU countries: Portugal and Cyprus. The SocialRobot project focuses more on providing care to those elderly people with light physical and cognitive disabilities, that are still capable to carry out their daily activities in their preferred environment. This project aims to the robotics field for solutions, offering functionality to support independent living, monitoring and maintaining safety and provide service and assistance in daily tasks. The SocialRobot will also provide solutions for stimulation and empowerment of socialization links, enhancing health and psychological wellbeing and reducing the care demand. The robotic design for the SocialRobot promotes a natural interaction for its users, thus it should have the ability to communicate in the most natural way possible. The integration of such communication, joined with behavior analysis and social adaptation, is one of the major challenges of the project.

1.4 Approach

A communication system should be considered as an independent module, capable of integration for any robotic system. The initial goal comprehended the implementation of computer vision modules, for the facial expression recognition in order to estimate human emotions, but as time went by we realized that the audio analysis was a field with more perceivable information and it became the main focus for this approach. Computer vision techniques are useful, and can fit in for future work, but usually we capture and express the most important features for communication processes trough our auditory and speech systems, thus our vision sense can be considered as a complement for those systems. The initial speech recognition module, was intended to be used trough a headset microphone, and that approach would imply a dependency of such device, which requires constant use, causing discommodity to the user. The use of only one microphone on a robotic platform, would probably lead us to unsatisfactory results and even create a spatial dependency. Those reasons, allied with the hardware resources, decided the use of a Microsoft Kinect [50]. The Kinect has a microphone array composed of four elements. This new setup brought the ability to perform sound source localization. Trough the sound source localization module it is possible to estimate the direction of a sound source, which can be useful, not only but also, for the localization of audio distress signals. Due to its advantages the microphone array is also used for the speech recognition module. Two approaches were explored in the speech recognition module implementation. In a preliminary stage we analyzed the captured audio from the four channels and use it as a single audio stream for the speech recognizer. This approach revealed some challenges for the speech recognition task, regarding a spatial dependency for the user and the communication system, resulting in low recognition rates for non frontal positions towards to the Kinect's microphone array. Those results lead to the second approach, that consists in the separated analysis for each sound wave captured by the microphones. Thus, it was possible to have four recognition hypothesis, and by relating them with the sound direction, the best recognition result can be chosen. This last approach revealed itself as an improvement, since it minimizes the initial spacial dependency for the speech recognition process.

In order to give social awareness to a robot, a emotion recognition module is also required. Depending on a person's emotional state, the contours of their sentences can be perceived with different meanings. Thus aligning the emotions with the sentences context will lead to a better understanding of human's real intentions. However, the emotion recognition by itself can also alert to abnormal behavior situations and corresponding social behaviors.

Voice emotion recognition can be perceived as a machine learning problem, thus it needs to pass through some stages until reaching a recognition output. The first stage realizes the audio data acquisition that will be segmented and forwarded into a feature extraction stage. The next stage is responsible for the training of a probabilistic model, containing the information about the features that characterize the recognizable emotions. Finally, the recognition of the emotions is in the classification stage, which analyses new audio data and performs a comparison with the model's learning data, estimating the most probable corresponding emotional state.

All the described techniques address important problems for human-robot interaction that can be considered independent from each other, but their assemble in a communication system, as in figure 1.1, enriches their potential into achieving more complex problems, such as social behavior analysis.

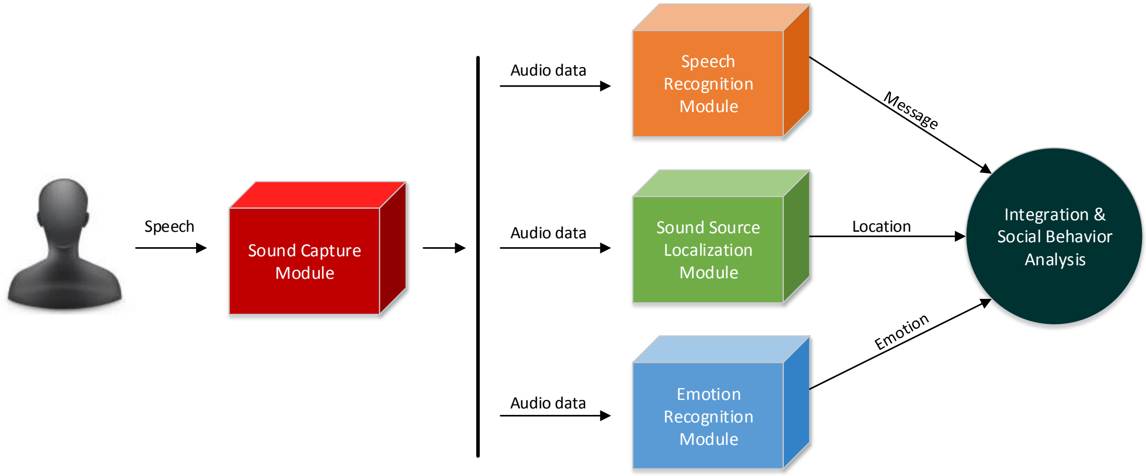

Speech
Sound Capture Module
Audio data
Audio data
Audio data
Speech Recognition Module
Sound Source Localization Module
Emotion Recognition Module
Message
Location
Emotion
Integration & Social Behavior Analysis

Figure 1.1: Module organization for integration and social behavior analysis.

1.5 Outline of the document

The first chapter introduces the context and motivation of this dissertation, the Social Robot project and the approach for the design concept of the communication system for social behavior analysis.

Chapter two reviews the related work and state of the toolkits for the sound source localization, speech recognition and emotion recognition modules considered for the integration, in order to reach a solution for the social behavior analysis.

In chapter three it is made a description of the use case scenarios, introducing the intended communication system functionality for the SocialRobot.

Chapter four is responsible for a more detailed explanation of the toolkits and their methods, also introducing Robot Operating System (ROS) and describing the implementation and integration of all developed ROS packages included in the system.

Chapter five goes into the details of the experimental setup and its results, regarding not only the accomplishment of the use case scenarios, but also a independent evaluation of the implemented modules.

A more extensive discussion of the system and its results is addressed in chapter six, opening the way for some future improvements and addictions, also described in that chapter.

Finally, the last chapter summarizes the work, providing final conclusions and prospects for this research.

Chapter 2

State of the Art

This chapter introduces, context and applicability for the modules in figure 1.1, also listing some related works and possible toolkits for implementation.

2.1 Sound Source Localization

The works and techniques described in this section, present an overview and guide lines for the implementation of sound source localization (SSL) systems. The toolkits research was crucial to choose the best suitable toolkit for the sound source localization module, see figure 1.1, reaching the adoption of the HARK toolkit. A brief overview of HARK's features is present in the toolkits subsection. Furthermore, a more detailed description is addressed in subsection 4.4.1.

2.1.1 Introduction

The integration of robots in human's social life requires communication capabilities, and one of the most important types of communication is the verbal communication, which supposes the ability to maintain a dialogue. In order to provide this ability to a robotic platform, it is necessary to implement an artificial auditory system that allows it to hear and recognize what is said by people. For speech recognition, it is often adopted a method to capture sound through headsets used by human users, but this method does not promote the expected interaction with the robot, since it entails the continuous use of such devices, which aren't suitable for real environments and patterns of communication between humans.

When humans communicate, they focus their attention in the speaker, so they need to know were the speaker is located. Most commonly the speaker is facing the person with whom he wants to communicate, but sometimes the speaker is not in the vision range of the listener and the communication tends to be inefficient until the listener focus

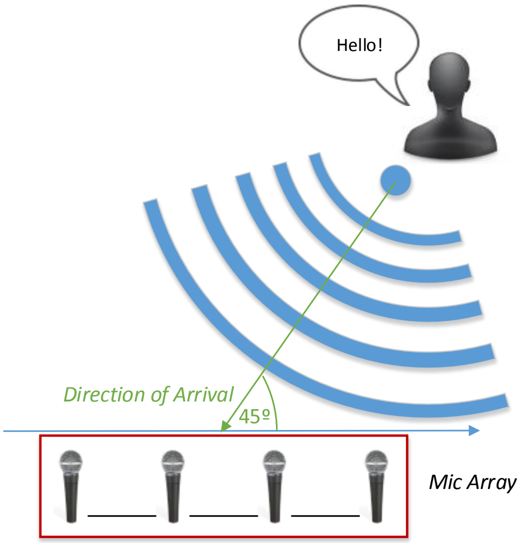


Figure 2.1: Sound Source Localization - Sound's direction of arrival.

his attention. Thus, in order to communicate effectively, the listener localizes the speech source, through his auditory system, and focus his attention towards the speaker. This is a problem that usually appears in real environments with multiple persons and simultaneous conversations, thus we should focus on the desired speech and abstract ourselves from the others. Human ears can correspond to two microphones in a robotic system, but human's auditory system can't be simplified as a sound input device. Humans have a complex auditory systems beyond the ears that send information to the most complex organ of them all, the brain. With the intention to integrate robots into human's real environments it is necessary to give them similar abilities. As it is not yet possible to recreate human intelligence we need to separate the robots auditory problems. Thus it is known that a robot has the ability to acquire sound through its microphones, but the problem remains in the separation and focus on the desired speech or speeches. Moreover, in order to solve those problems it is necessary to address and deliver sound source localization, see figure 2.1, and sound source separation abilities to the robots sound acquisition systems. The combination of this techniques is usually presented to enhance speech signals and improve speech recognition results.

2.1.2 Related Work

Ben Rudzyn et al. described the implementation of a novel real time robot audition system which combines a 3D sound localization system and a voice characterization (VC) system [36]. The localization system employs a four microphone array and uses the time delay estimation method. The system accuracy is improved through the use of a correlation confidence threshold and a median filter. The inclusion of a sound localization system identifies the spatial coordinates of a sound source relative to the robot in three dimensions: azimuth (θ), elevation (ϕ) and distance (ρ). The sound localization system, based on time delay estimation, assumes that the propagation of a sound wave over the different path lengths will result in a delay of arrival between the microphones. To measure the time delay they used weighted cross correlation (WCC) function. Combining the time delay information with the known geometry of the microphone array they obtain the direction of arrival (DOA) estimate for the sound source. Preceding work already introduced microphone arrays for speaker localization.

Takeshi Yamada et al. proposed a robust speech recognition with Speaker Localization by an Arrayed Microphone (SLAM) to realize hands-free speech interface in noisy environments [48]. In order to localize a speaker direction accurately in low SNR conditions, it was introduced a speaker localization algorithm based on extracting a pitch harmonics. The SLAM system is composed of a speaker localizer, a speech enhancer and a speech recognizer. The localizer can detect the speaker's direction under the existence of several noise sources, from which the speech enhancer obtains an enhanced speech signal. The delay-and-sum beamformer is used as a microphone array signal processing. The speaker localization algorithm consists in a frequency analyzer in order to apply the delay-and-sum beamformer for broadband signals and a Sound source direction estimator in order to estimate sound source directions. Finally the speaker direction is detected from among the sound source directions estimated.

A new talker localization algorithm, consisting of two algorithms, was proposed by Takanobu Nishiura and Satoshi Nakamura. One is DOA (Direction of Arrival) estimation algorithm for multiple sound source localization based on CSP (Cross-power Spectrum Phase) coefficient addition method. The other is statistical sound source identification algorithm based on GMM (Gaussian Mixture Model) for localizing the target talker position among localized multiple sound sources [30]. The talker localization algorithm estimates multiple sound DOAs with the CSP coefficient addition method after multiple sound signals are captured. Then, sound signals of estimated DOA are enhanced by steering the microphone array to them. Finally, the talker can he localized after identification between "speech" or "non-speech" using statistical speech and environmental sound models among the enhanced multiple sound signals. The system recognizes the input from a sound source identified as being "speech".

A more concrete example for sound source localization was given by Cha Zhang et al. [49] introducing microphone arrays in distributed meetings. Their approach was based in a unified maximum likelihood (ML) framework to capture superior speech sound and perform speaker localization. The proposed method is closely related to steered response power-based algorithms. They demonstrated within the ML framework how reverberation can be dealt with by introducing an additional term during noise modeling, and how the unknown directional patterns of microphone gains can be compensated for from the received signal and the noise model. This work resulted in a new and efficient SSL algorithm that can be applied to various kinds of microphone arrays, even for challenging cases such as circular directional arrays with unknown directional patterns. Additionally, this ML derivation demonstrates that the traditional minimum variance distortionless response (MVDR) beamforming technique is equivalent to the ML-SSL.

Other approaches can also be taken into consideration, such as the introduction of Bayesian inference for the sound source estimation. Cátia Pinho et al. presented a Bayesian system of auditory localization in distance, azimuth and elevation using binaural cues only [31]. The binaural system is also integrated in a spatial representation framework for multimodal perception of 3D structure and motion, the Bayesian Volumetric Map (BVM). The Bayesian binaural system is composed of a monaural cochlear unit, which processes the pair of monaural signals coming from the binaural audio transducer system by simulating the human cochlea, so as to achieve a tonotopic representation (i.e. a frequency band decomposition) of the left and right audio streams; the binaural unit, which correlates these signals and consequently estimates the binaural cues and segments each sound-source; and, finally , the Bayesian 3D sound-source localization unit, which applies a Bayesian sensor model so as to perform localization of sound-sources in 3D space. This work contributed with a novel probabilistic solution that produces localization estimation based on binaural cues yielded by a robust binaural processing unit.

Another probabilistic approach proposed by Volker Willert et al., introducing a biologically inspired and technically implemented sound localization system to estimate the position of a sound source in the frontal azimuthal half-plane [45]. The process is based in the extraction of binaural cues, using cochleagrams generated by a cochlear model that serve as input to the system. The basic idea of the model is to separately measure inter-aural time differences and inter-aural level differences for a number of frequencies and process these measurements as a whole. This leads to two-dimensional frequency versus time-delay representations of binaural cues, so called activity maps. A probabilistic evaluation was presented to estimate the position of a sound source over time based on these activity maps. Learned reference maps for different azimuthal positions are integrated into the computation to gain time-dependent discrete conditional probabilities. Those probabilities were combined over frequencies and binaural cues to estimate the sound source position. The binaural cues used in this approach are based on differences in timing and level of the sound at each of the two ears called inter-aural time differences (ITDs) and inter-aural level differences (ILDs).

Using a larger set of microphones, Yuki Tamai et al. described a “32-channel circular microphone array” which can localize individual sounds from many sound sources in 360 degrees [41]. The microphone array was used as the robot’s audition system. Because the microphone array is composed with many microphones, it has the tolerance for not only the environmental noise, but also the one caused from the robots. Moreover, it can localize and separate sound sources in the omni-direction. In their research, sound localization is achieved by Delay and Sum Beam Forming algorithm (DSBF). The DSBF method is a technique for forming a strong direction characteristic in the direction of the purpose by aligning all time shifts and amplitudes of sound waves inputted from each microphone and adding.

By taking into account the directivities of the source and microphones as well as the source-microphone distances, Bob Mungamuru and Parham Aarabi presented enhanced sound localization algorithms [27], offering improved sound localization accuracy and estimates of the source orientation. Two examples of such algorithms were developed. The temporal ML algorithm, using a simple time-domain model for the speech source and the weighted SRP-PHAT algorithm that modifies the temporal ML in an attempt to mitigate the adverse effects of reverberations. The common thread shared between those algorithms, is that the search space extends over three dimensions instead of two cartesian coordinates, by including the speech orientation. As a result, both the magnitude differences and phase differences between the signals observed at microphones are used as discriminative information.

2.1.3 Toolkits

The interest for SSL and its research, resulted in the creation of some toolkits. Those toolkits are intended to save time in the implementation of SSL capabilities into robotic systems.

The ManyEars toolkit [18] presents a robust sound source localization and tracking method using an array of eight microphones. The method is based on a frequency-domain implementation of a steered beamformer along with a particle filter-based tracking algorithm. Tests on a mobile robot show that the algorithm can localize and track in real-time multiple moving sources of different types over a range of 7 meters. ManyEars is implemented in C as a modular library, with no dependence on external libraries. The source code is available online under the GNU GPL license. A Graphical User Interface (GUI) is used to display in real time the tracked sound sources and to facilitate configuration and tuning of the parameters of the ManyEars library. ManyEars needs at least four microphones to operate, and the number of microphones used influences the number of sources that can be processed. It has mostly been used with arrays of eight microphones, to match the maximum number of analog input channels on the sound cards used. The ManyEars Library is composed of five modules: Pre-processing, Localization, Tracking, Separation and Post-processing. These modules receive inputs and generate data using the microphones, potential sources, tracked sources, separated sources and post-filtered sources data structures. It is compatible with ROS and provides an easy-to-use GUI for tuning parameters and visualizing the results in real time. For more detailed information see [18].

HARK is open-sourced robot audition software and it consists in a set of modules to perform many auditory tasks [28], [29]. Such as audio signal input, sound source localization, sound source separation, acoustic feature extraction, and speech recognition. HARK does not provide a speech recognizer but integrates Julius as an external package to perform

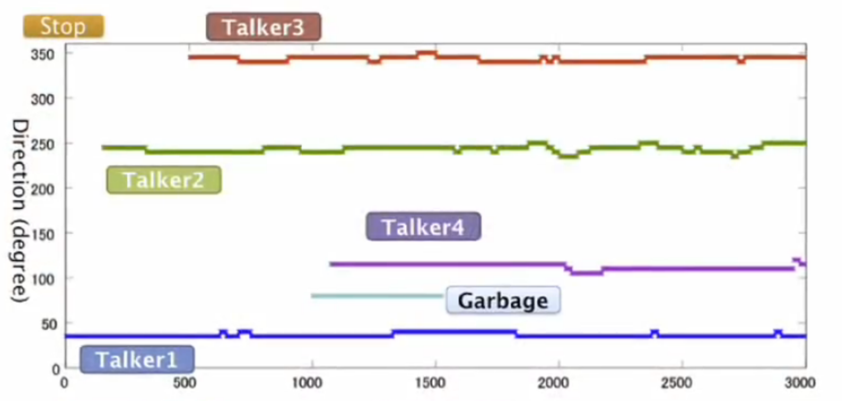


Figure 2.2: HARK sound source localization for four simultaneous talkers in a 360º range.

the speech recognition on the separated sound sources. All the modules for the auditory tasks are implemented in FlowDesigner which comes with the HARK software and is a free middleware equipped with a data flow-oriented GUI development environment that realizes high speed and lightweight module integration. In FlowDesigner, each module is realized as a C++ class.

Singe channel microphones are frequently used to capture continuous audio streams, but in order to gather more information and perform sound localization, they have been replaced by microphone arrays, which bring the ability to capture and retrieving multichannel audio data. HARK supports those both types of devices and has a Microsoft Kinect driver which allows the use of its microphone array as an ALSA based device.

HARK is capable of performing sound source localization with microphone arrays through its localizeMUSIC module, which uses the MUltiple SIgnal Classification (MUSIC) method. The MUSIC method localizes sound sources based on source positions and impulse responses (transfer function) between each microphone of a microphone array. Impulse responses can be obtained by actual measurements or calculation with geometric positions of microphones, using HARK's harktool4[1]. The most recent version of HARK (v1.2.0) comes with GEVD-MUSIC and GSVD-MUSIC which are extended version of MUSIC and can suppress or whiten a known high power noise such as robot ego-noise and localize desired sounds under this noise conditions. The localization algorithm was also extended to localize sound sources in a 3D way. An example of HARK's sound source localization is represented in figure 2.2 and more functionalities can be discovered in [6].

2.2 Speech Recognition

This section introduces speech recognition and its related works and followed approaches. A list of toolkits is also presented, aiming to the fulfillment of the speech recognition module in figure 1.1.

[1] HARK tool for the creation of transfer function files

From all the described toolkits, we chose the CMU Sphinx for the speech recognition module implementation. A more detailed overview of what goes on inside this toolkit is presented in subsection 4.5.1.

2.2.1 Introduction

Nowadays the invitation to humans social lives has been extended to machines and consequently to robots. In order to robots have a active part in human society, they need to learn and improve ways to communicate and interact with people. One of the main goals and concerns for human-robot communication is presented in the dialog, and for that purpose many researches over the years have been done in order to make machines able to perform speech recognition. In general, there are two approaches to speech recognition, the acoustic-phonetic approach and the pattern recognition-based approach [42]. The first approach relies on the segmentation of continuous speech into well-defined regions, from which the assignment of phonetic labels is based on measured properties of the speech features from the speech signals. On the other hand, the pattern recognition-based approach, models the basic speech units through a lexical description of words in the vocabulary.

Almost all speech recognition systems perform the analysis of speech utterances through parametric representations of their waveforms. From the analyses of the speech utterances, feature vectors are extracted, which serve as input to classification modules that estimate linguistic units like words and phonemes, as described in figure 2.3. Moreover, sentences are formed by the concatenation of words, according to syntax and semantic rules. To accomplish this task, acoustic, grammar and language models are combined to provide a more accurate mechanism for estimating the probability of words in an utterance given preceding words, context and other speech properties.

2.2.2 Related work

Speech recognition applications have been developed over the years. These applications rely on the ability to recognize words and sentences present in speech utterances, through statistical methods. Words can be perceived as a sequence of sounds, usually addressed as phonemes. That transition brings a time relationship between phonemes in order to express a certain word. To include that time dependency, in the recognition processes, the common approach relies on Hidden Markov Models. Lawrence Rabiner was one of the initial spreaders of the HMM, introducing the theory and ways to use it for speech recognition applications [35].

With the evolution in this field, many other problems and solutions have been addressed. Gales and Young concerned their approach to noise robustness, using parallel models combination (PMC) [15]. The basic concept behind PMC is that the performance of

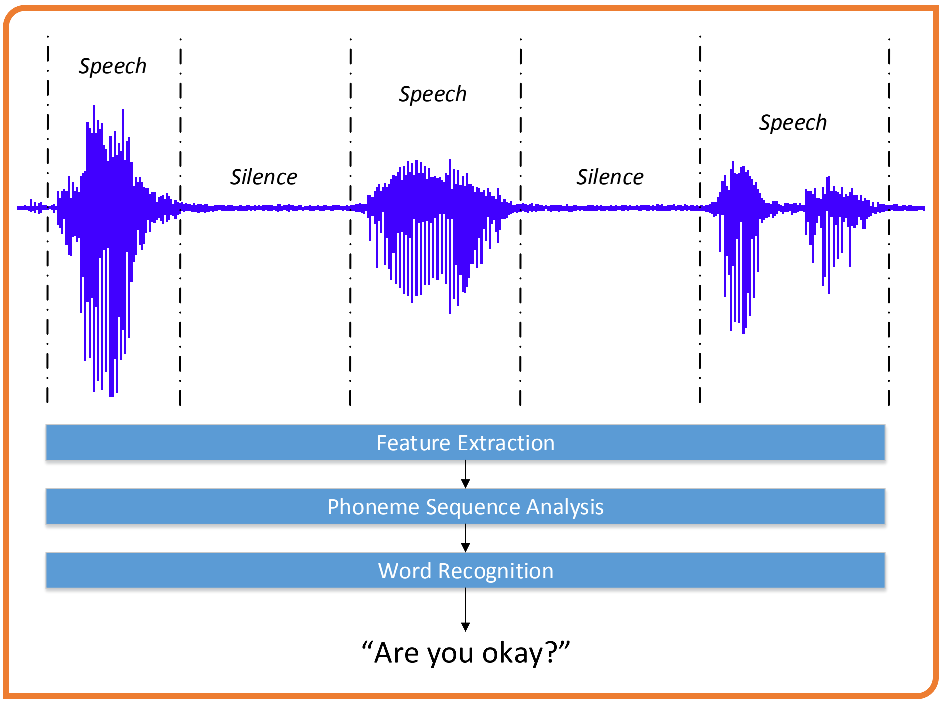


Figure 2.3: Recognition of a speech utterance.

speech recognition systems is optimal when there is no mismatch between training and test conditions. Invariably in real applications there is some mismatch, either form additive noise, or variations in the channel conditions. After training on clean speech data, the performance of the recognizer was found to be severely degraded when noise was added to the speech signal. Thus, they used the PMC method to achieved a compensation of the parameters in a computationally efficient manner. The PCM model used a standard HMM with gaussian output probability distributions, which restored speech recognition performance to a level comparable with that obtained when training directly in the noise corrupted environment.

Nathalie Virag focused on single channel speech enhancement. The approach was based on the introduction of an auditory model in a subtractive-type enhancement process, which attempts to estimate the short-time spectral magnitude of speech by subtracting a noise estimation from the noisy speech [44]. The developed algorithm was based on masking properties of the human auditory system, modeled by calculating a noise masking threshold, based on the assumption that the human listener tolerates additive noise as long as it remains below the threshold. That threshold allowed an automatic adaptation in time and frequency of the parametric enhancement system, finding the best tradeoff based on a criterion correlated with perception. This approach led to the reduction of the residual noise in the speech signals and improved the performance of the speech recognition task.

Sue Harding et al. described a perceptually motivated computational auditory scene analysis (CASA) system that combines sound separation according to spatial location with the missing data approach for robust speech recognition in noise [20]. Missing data time-frequency masks were created using probability distributions based on estimates of interaural time and level differences for mixed utterances in reverberated conditions. Those masks indicate which regions of the spectrum constitute reliable evidence of the target speech signal. The CASA system, exploits spatial location cues in order to improve the robustness of the speech recognition system in multisource, reverberant environments. Their approach, consists of two processing stages. In the first stage, acoustic features and binaural cues (ITD and ILD) are derived from an auditory model. The binaural cues are used to estimate a time–frequency mask, in which each element indicates whether the corresponding acoustic feature constitutes reliable evidence of the target speech signal. In the second stage, the acoustic features and the time-frequency mask are passed to a missing data speech recognition system, which treats reliable and unreliable features differently during decoding.

Later, Liang-Che Sun and Lin-Shan Lee proposed novel approaches for equalizing the modulation spectrum for robust feature extraction in speech recognition [40]. They used histogram equalization (HEQ) to reduce the mismatch between speech feature parameters from the training set and the noise corrupted testing sets. The histogram equalization of the modulation spectrum was performed for each speech utterance with reference to the histogram obtained from clean training data, or by equalization with two sub-bands on the modulation spectrum. For the magnitude ratio equalization, they also considered each speech utterance separately, defining for each one the respective magnitude ratio of lower to higher modulation frequency components. The equalization was, once again, referenced with the value obtained from clean training data. Their approaches are described as temporal filters that are adapted to each testing utterance. They also shown that additional improvements could be obtained by integrating cepstral mean and variance normalization (CMVN), histogram equalization (HEQ), higher order cepstral moment normalization (HOCMN), or the advanced front-end (AFE).

Another approach, presented by Alberto Sanchis et al. in 2012, proposes a word-based Naive Bayes classification model for confidence estimation in speech recognition [37]. The model has a combination of word-dependent and word-independent Naive Bayes models. Their classification model is empirically compared with confidence estimation based on posterior probabilities computed on word graphs. The computation of word graph-based posterior probabilities relies on a word graph aligned with a recognized sentence. Each internal word graph node corresponds to a recognition time frame in which a transition between words has been produced in the search process, and each word is labeled with its corresponding posterior probability. The confidence estimation can be seen as a conventional pattern classification problem in which a set of features is obtained for each hypothesized word in order to classify it as either correct or incorrect recognition result. The recognized sentence is encountered by choosing the hypothesized words with the higher posterior probabilities along the word graph.

2.2.3 Toolkits

Since speech recognition applications have been developed for many years, achieving high rates of accuracy, most of the research was turned into the enhancement of these recognition systems. To abstract new researchers and developers from the modeling problems and basic concepts of speech recognizers, some toolkits have emerged.

One of the most known toolkits is the Hidden Markov Model toolkit (HTK). HTK is a portable software toolkit for building and manipulating Hidden Markov Models [47], developed by the Cambridge University Speech Group. HTK consists of a set of library modules and tools that provide sophisticated facilities for speech analysis, HMM training, testing and results analysis. The software supports HMMs using both continuous density mixture Gaussians and discrete distributions and can be used to build complex HMM systems [2]. HTK also presents an application to facilitate building experimental applications, the ATK[2], which allows the compilation and test of novel recognizers, built using customized versions of HTK.

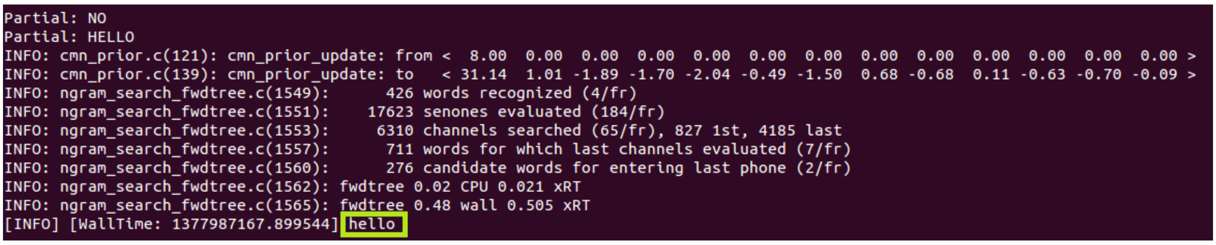
Kaldi is similar in aims and scope to the HTK mentioned before, and it is also written in C++. The main intended use of Kaldi is for acoustic modeling research. Kaldi is an open-source toolkit, based on finite-state transducers, with detailed documentation and scripts for building complete speech recognition systems [32]. Kaldi's core library supports modeling of arbitrary phonetic-context sizes, acoustic modeling with standard Gaussian mixture models (GMM) and subspace Gaussian mixture models (SGMM).

Kyoto University also contributed with a two-pass large vocabulary continuous speech recognition decoder software for speech-related researchers and developers, named Julius, which is capable of almost real-time decoding in 20k word dictation tasks [25]. Julius incorporates major search techniques, such as tree lexicon, N-gram factoring, cross-word context dependency handling, enveloped beam search, Gaussian pruning and Gaussian selection. It also supports various HMM types, such as shared-state triphones and tied-mixture models, with any number of mixtures, states, or phones.

Within the most interesting toolkits, there is the Carnidge Melon University's CMU Sphinx [1]. It is organized as different packages of tools for different applications. The CMUSphinx tools provide acoustic model and language model training, as well as recognizer libraries for speech recognition applications. This toolkit as evolved through the years, and some speech recognizers have been developed and enhanced along the way. Pocketsphinx is one of those recognizers, which recently has been considered to be included in Ubuntu's[3] latest versions. This recognizer is also known for its use in hand-held devices, with improved performance for acoustic feature calculation, Gaussian computation, Gaussian mixture model computation, and HMM evaluation [22]. Furthermore, the

[2] Real-Time API for HTK

[3] Open-source Linux distribution - http://www.ubuntu.com



```
Partial: NO
Partial: HELLO
INFO: cmn_prior.c(121): cmn_prior_update: from <  8.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00 >
INFO: cmn_prior.c(139): cmn_prior_update: to   < 31.14  1.01 -1.89 -1.70 -2.04 -0.49 -1.50  0.68 -0.68  0.11 -0.63 -0.70 -0.09 >
INFO: ngram_search_fwdtree.c(1549):      426 words recognized (4/fr)
INFO: ngram_search_fwdtree.c(1551):    17623 senones evaluated (184/fr)
INFO: ngram_search_fwdtree.c(1553):     6310 channels searched (65/fr), 827 1st, 4185 last
INFO: ngram_search_fwdtree.c(1557):      711 words for which last channels evaluated (7/fr)
INFO: ngram_search_fwdtree.c(1560):      276 candidate words for entering last phone (2/fr)
INFO: ngram_search_fwdtree.c(1562): fwdtree 0.02 CPU 0.021 xRT
INFO: ngram_search_fwdtree.c(1565): fwdtree 0.48 wall 0.505 xRT
[INFO] [WallTime: 1377987167.899544] hello
```

Figure 2.4: Pocketsphinx ROS package recognizing the "hello" word from a live speech signal.

Pocketsphinx recognizer has its own ready-to-use ROS package. A simple word recognition example can be realized in figure 2.4.

2.3 Voice Emotion Recognition

Introducing the advances and prospects into emotion recognition for robotic systems, this section exposes related works and its approaches. Some toolkits for the emotion recognition module, in figure1.1, are also described. In advance, the openEAR toolkit was the one adopted for the module implementation and a more detailed description can be viewed in subsection 4.6.1.

2.3.1 Introduction

To insert listening capabilities in robotic system, it is common to use microphones, but this addition of hardware isn't enough for robots to understand what their are listening to. Since the microphones only provide signal output, it is necessary to add intelligence to robots in order for them to analyze these signals and extract useful information.

Over the years many developments have been done and many speech recognition systems were implemented, with the capability of translating the speech signals from the microphones into words. More recently the humans emotions also became a very interesting research field, in order to add more information about the human's social behavior. Usually complemented with a speech recognition system, the emotion recognition systems can help the robot to understand human's intentions and adapt their own behavior according to human's emotional states.

With the increasing popularity of robots in human's society, there have been some projects like the Social Robot aiming to introduce elderly assistive robots. These robots are not expected to replace human caregivers, but they are expected to help the caregivers and interact directly with the elderly population. In order to perform a reliable interaction the robots will need to have human similar capabilities to communicate and adapt their response according to the the humans behavior. This means that the robots need to have a speech recognition system and a emotion recognition system.

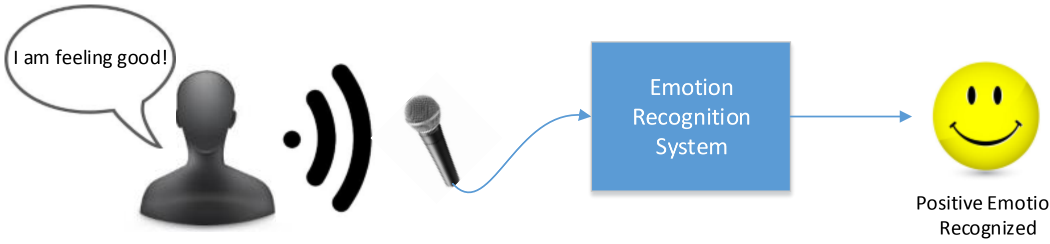


Figure 2.5: Emotion recognition from audio input.

The emotion recognition system is typically implemented through image processing, analyzing the human's facial expressions. This approach has been followed in previous works proving good results for the emotion recognition problem. Only analyzing the human's facial expressions, brings a dependency of location between the human and the robot, since the human needs to be in the vision range of the robot in order to capture the human's facial expressions. Recognizing human's emotion through speech signals has also been an interesting challenge in robotics. This approach doesn't rely on the location dependency, since the sound can be acquired through any location of the speaker, i.e. even if the speaker isn't in the vision range of the robot, the microphone will capture his speech signal, providing the robot a input for the speech and emotion recognition systems. Figure 2.5, represents an overview of the emotion recognition setup from emotional speech input.

2.3.2 Related Work

Emotional speech has specific features, according to the emotional state, thus a robot needs to a have the ability to recognize them in order to infer the actual emotion in the speech. This brings the necessity for feature acquisition and characterization to perform emotion classification, facing a problem of machine learning for the emotion recognition.

Lee et al. used various acoustic feature sets and classification algorithms for classifying spoken utterances based on the emotional state of the speaker [26]. For the classification process, they used a linear discriminant classifier (LDC), k-nearest neighborhood (k-NN) classifier, and a support vector machine (SVM) classifier. The classification consisted in the recognition of two emotion classes: negative and non-negative. They used two feature sets, one obtained from the utterance-level statistics of the pitch and energy of the speech, and another feature by principal component analysis (PCA). The PCA was used to to discover, and reduce, the underlying dimensions of the feature space. They used ten utterance-level statistics derived from $F0$ and energy as the acoustic features for emotion recognition, and their results noticed that SVC outperforms LDC in the light of generalization, due to the data sparsity encountered in the real world applications.

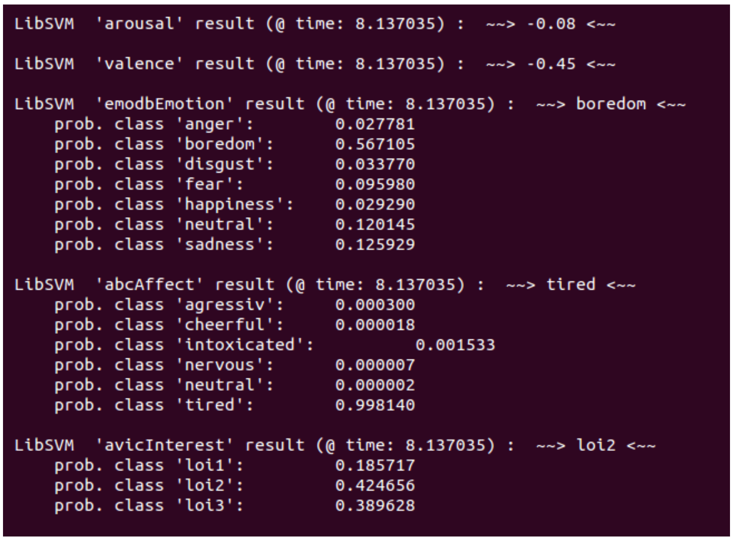
Chuang and Wu presented an approach to emotion recognition from speech signals and textual content [12]. The proposed emotion recognition system is intended to classify six basic emotions, including happiness. sadness, anger, fear, surprise and disgust. If the emotion intensity value of a recognized emotion is lower than a predefined threshold, the emotion output will be neutral. Their system can detect the emotion from two different types of information: acoustic signal and textual content. For the acoustic module, an initial acoustic feature set that contains totally 33 features was firstly extracted and adopted. These acoustic features contain several possible aspects, such as intonation, timbre, acoustics, tempo, and rhythm. They also extract some features to represent the special intonations, such as trembling speech, unvoiced speech, and crying speech. Among these diverse features, the most significant features are selected by the principle component analysis (PCA) to form an acoustic feature vector. The acoustic feature vector is used as input to a Support Vector Machine classifier (SVM) to determine the emotion output. For the textual emotion recognition module, it was assumed that the emotion reaction of an input sentence is essentially represented by its word appearance. Two primary word types, "emotional keywords" and "emotion modification words", are manually defined and used to extract emotion from the input sentence. All of the extracted emotional keywords and emotion modification words have their corresponding "emotion intensity values" and "emotion modification values." which are manually defined. For each input sentence, the emotion intensity values are averaged and triggered by the emotion modification values to give the current emotion output. The final emotion output comes from the combination of the textual and the acoustic emotion modules with the emotion history.

Feature selection is a important step towards emotion recognition. Koolagudi and Rao, analyzed different speech features and their combination for emotion recognition purposes [23]. In their work, the selected speech features extracted are: excitation source, spectral and prosodic features. Auto associative neural networks (AANN) were used to capture the emotion specific information from excitation source features, Gaussian mixture models (GMM) for developing the models using spectral features, and support vector machines (SVM) to discriminate the emotions using prosodic features. From those features, different combinations were addressed and tested for emotion recognition. Their results conclude that each of the proposed speech features has contribution towards emotion recognition, while their combination improved the emotion recognition performance, indicating the complementary nature of the features.

In order to overcome the lack of suitable training data for speech emotion recognition, Schuller et al. introduced synthesized speech for model training and combination with human databases [39]. The acoustic features extraction was performed with the open-EAR toolkit, and for the classification they chose a linear kernel Support Vector Machine (SVM). The performance of the different combinations of human and synthesized speech in binary arousal and valence classification of eight popular human speech databases, demonstrated that combining human and synthesized speech increases the expected performance while decreasing the performance variability caused by training with human speech databases. In many cases, the training only on synthesized speech, shown to be competitive against training on human speech databases. As the quantity and quality of human emotional speech databases is still reduced, the usage of synthesized speech is a valid option, also providing ways to improve speech emotion recognition systems.

2.3.3 Toolkits

There are also some existing toolkits regarding emotion recognition problems. PRAAT [9] is a toolkit for speech researchers, featuring speech analysis, labeling and segmentation tools and some learning algorithms. Although this toolkit many functionalities, it seams to be left behind by its opponents. Other toolkits, integrate more machine learning algorithms and functionalities. The WEKA toolkit supports data preprocessing, clustering, classification, regression, visualization, and feature selection [19]. However, all of WEKA's techniques are predicated on the assumption that the data is available as a single flat file or relation. Another toolkit is the openEAR [14], which combines audio recording, feature extraction, and classification to evaluation of results, and pre-trained models while being very fast and highly efficient. All feature extractor components are written in C++ and can be used as a library, facilitating integration into custom applications. Also, openEAR can be used as an out-of-the-box emotion live affect recognizer, see figure 2.6, for various domains, using pre-trained models which are included in the distribution.



```
LibSVM  'arousal' result (@ time: 8.137035) :  ~~> -0.08 <~~

LibSVM  'valence' result (@ time: 8.137035) :  ~~> -0.45 <~~

LibSVM  'emodbEmotion' result (@ time: 8.137035) :  ~~> boredom <~~
    prob. class 'anger':        0.027781
    prob. class 'boredom':      0.567105
    prob. class 'disgust':      0.033770
    prob. class 'fear':         0.095980
    prob. class 'happiness':    0.029290
    prob. class 'neutral':      0.120145
    prob. class 'sadness':      0.125929

LibSVM  'abcAffect' result (@ time: 8.137035) :  ~~> tired <~~
    prob. class 'agressiv':     0.000300
    prob. class 'cheerful':     0.000018
    prob. class 'intoxicated':          0.001533
    prob. class 'nervous':      0.000007
    prob. class 'neutral':      0.000002
    prob. class 'tired':        0.998140

LibSVM  'avicInterest' result (@ time: 8.137035) :  ~~> loi2 <~~
    prob. class 'loi1':         0.185717
    prob. class 'loi2':         0.424656
    prob. class 'loi3':         0.389628
```

Figure 2.6: openEAR live emotion recognition from speech signals using the pre-trained models and LibSVM library.

Chapter 3

Use Case Scenarios

3.1 Introduction

The addressed use case scenarios in this chapter are based in some scenarios presented for the SocialRobot project [34], assuming the integration of the sound source localization, speech recognition and emotion recognition modules intended for this thesis. The scenarios consider hypothetical situations and the respective response of the SocialRobot. Some features and abilities for the SocialRobot are presumed, such as navigation and a knowledge database, among others, but the description of those use cases intends to demonstrate and validate the need and applicability for the integration of the modules considered in this thesis.

A detailed description of the three use case scenarios considered is given in the sections 3.2, 3.3 and 3.4. The hypothetical situations are introduced by the figures 3.1, 3.3, 3.5 and an overview of the steps of each use case can be realized from figures 3.2, 3.4 and 3.6 respectively.

3.2 Use Case Scenario 1

Fall Detection and Alerting - Home Use

Ana is a 76 year old woman with light cognitive and physical disabilities. Ana has regular support from her son John. As the years gone by she shown some light mobility problems.

Since John is at work most of the time and cannot take continuous care of his mother, he found the SocialRobot solution very convenient, which made him feel more comfortable to leave his mother alone at home. The SocialRobot usually stays in the living room, which is the most used zone of the house by Ana. One day Ana was walking towards the kitchen when she lost her balance and fell down. Ana shouted “Help” and the SocialRobot recognized the help call through its speech recognition system, and realized that the sound

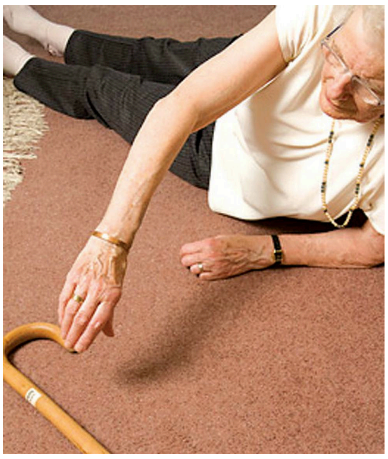


Figure 3.1: Elderly woman lying on the floor

was coming from the kitchen through its sound source localization system, which led the SocialRobot to Ana's location. Already near Ana the SocialRobot realizes that Ana is lying on the floor and asks her "Do you want me to request help?", Ana answers "Yes" and the SocialRobot recognizes the distress characteristics in her voice through its speech and emotion recognition systems. The SocialRobot told her to stay calm and that help was coming soon, immediately the SocialRobot alerted her son John and Ana's very good neighbors about the incident and asked for their support. SocialRobot connected Ana, using a Skype video call, with her son John. Seeing and talking to each other and also knowing that help was coming, made them both feel more comfortable.

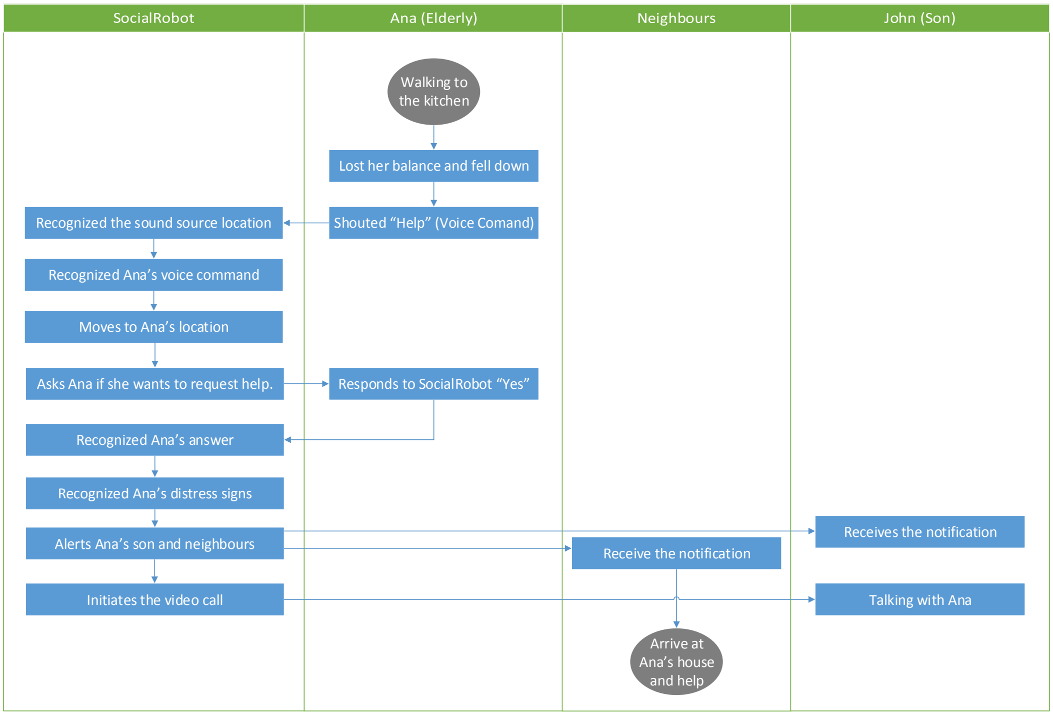

SocialRobot
Ana (Elderly)
Neighbours
John (Son)
Walking to the kitchen
Lost her balance and fell down
Shouted "Help" (Voice Comand)
Recognized the sound source location
Recognized Ana's voice command
Moves to Ana's location
Asks Ana if she wants to request help.
Responds to SocialRobot "Yes"
Recognized Ana's answer
Recognized Ana's distress signs
Alerts Ana's son and neighbours
Receives the notification
Receive the notification
Initiates the video call
Talking with Ana
Arrive at Ana's house and help

Figure 3.2: Flowchart of the Use Case Scenario 1

3.3 Use Case Scenario 2

Recognizing Abnormal Behavior and Alerting - Elderly Care Centre Use

Figure 3.3: Elderly in bed, assisted by the centre staff called by the SocialRobot

Silvia is a 75 years old woman with light cognitive and physical disabilities. Besides the care provided by the elderly care centre's staff, Silvia also enjoys the presence of the SocialRobot that takes continuous care of her since she arrived at the elderly centre.

One morning the SocialRobot goes to Silvia's bedroom to wake her up, as usual, but Silvia insisted to stay in bed. The SocialRobot approached her and asked if she was feeling ok, Ana responded "No". Knowing that Ana as stomach problems, it asked "Is it your stomach?" and she answered "Yes", immediately the SocialRobot recognized what Silvia said and the distress characteristic in her voice through its speech recognition and voice emotion detector systems and alerted the care centre's staff about the situation, performing a Skype video call and sending a message with details. The elderly centre's nurse came to Silvia's bedroom with her pills for the stomach pain. After a few minutes Silvia's pain was starting to disappear and she got up and went to breakfast. Silvia and the care centre staff were pleased with the SocialRobot that was very useful improving Silvia's assistance.

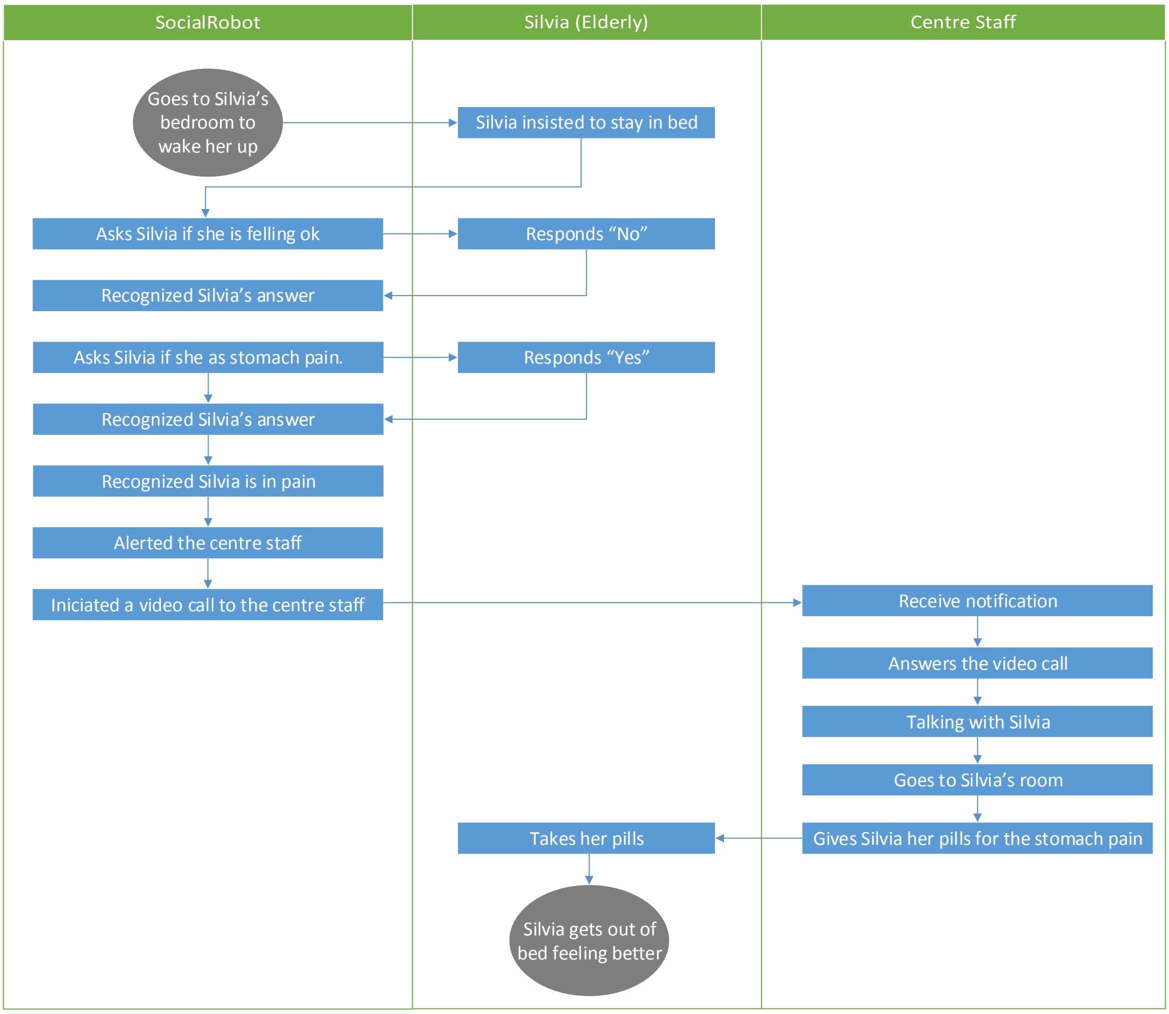

SocialRobot
Silvia (Elderly)
Centre Staff
Goes to Silvia's bedroom to wake her up
Silvia insisted to stay in bed
Asks Silvia if she is felling ok
Responds "No"
Recognized Silvia's answer
Asks Silvia if she as stomach pain.
Responds "Yes"
Recognized Silvia's answer
Recognized Silvia is in pain
Alerted the centre staff
Iniciated a video call to the centre staff
Receive notification
Answers the video call
Talking with Silvia
Goes to Silvia's room
Gives Silvia her pills for the stomach pain
Takes her pills
Silvia gets out of bed feeling better

Figure 3.4: Flowchart of the Use Case Scenario 2

3.4 Use Case Scenario 3

Service and Assistance - Elderly Care Centre Use

Kate is a 74 years old woman with some mobility problems. Kate also has a prescription for cholesterol and she needs to take her pills after her meals.

One day, after lunch, Kate went to the living room with her friends from the elderly centre and as their conversation went by she realized it was time to take her medicine, but she didn't have any pills with her.

Figure 3.5: Left side - Elderly socializing on the centre living room. Right side - Medication brought by the Social Robot.

The elderly care centre recently acquired the SocialRobot and it was well accepted and appreciated by the centre users. The SocialRobot was in the same living room as Kate and by asking it to approach her with a voice command, the SocialRobot recognized her location and went towards her direction. Already in the presence of the SocialRobot, Kate said "Get my pills" and the SocialRobot recognized Kate's request. The SocialRobot also recognized Kate's identity and knowing that she as cholesterol problems, immediately went to the care centre staff room and requested Kate's medication. The care centre staff picked up the correspondent pills and filled up a bottle of water and putted the items in the SocialRobot carrying compartment. The SocialRobot returned to Kate's location and she took her medicine. Kate thanked the SocialRobot, since it would be inconvenient for her to go to the staff's room due to her mobility difficulties.

Without the SocialRobot assistance Kate wouldn't probably take her medicine and maybe encounter health problems during the rest of the day.

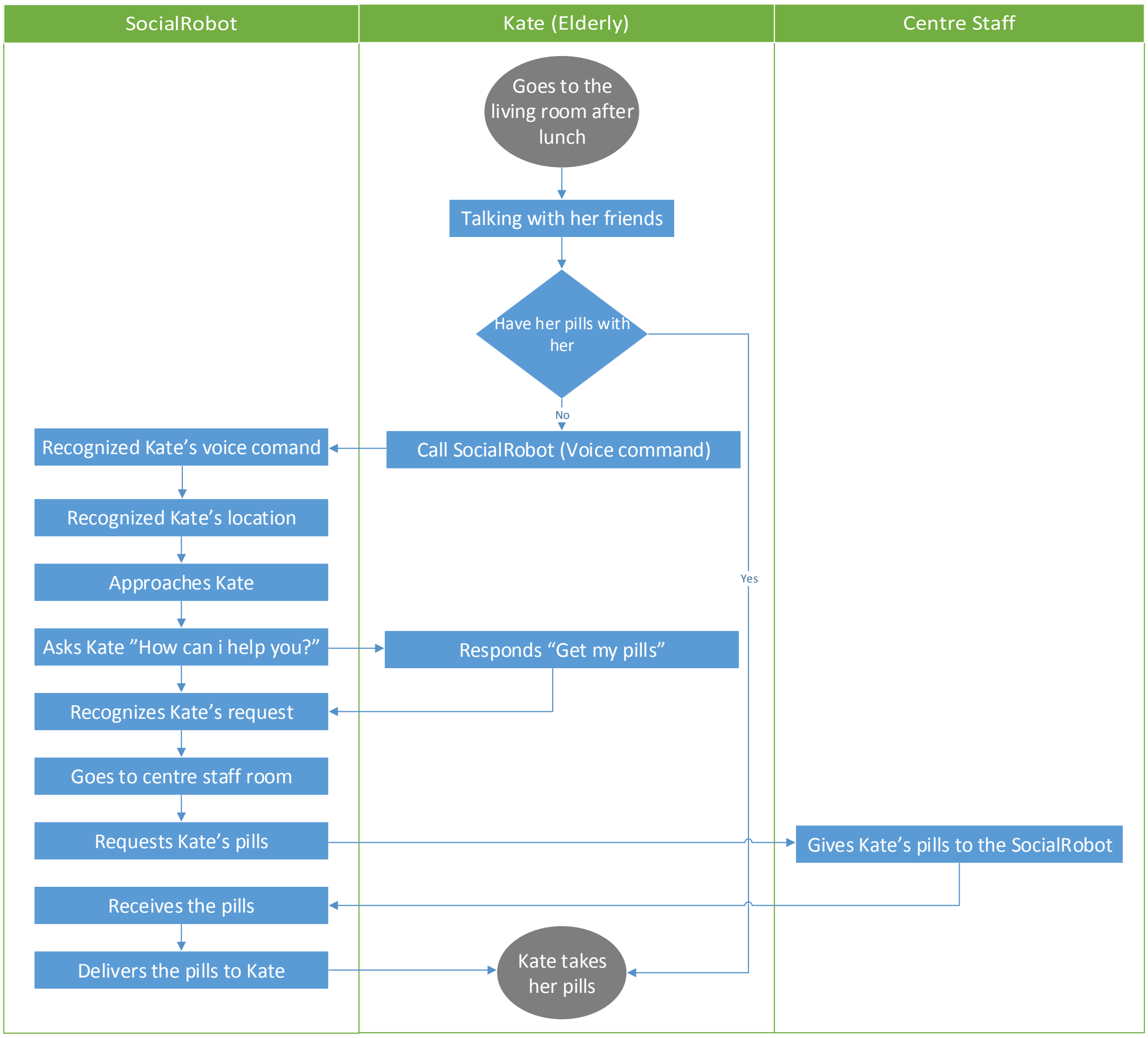

SocialRobot
Kate (Elderly)
Centre Staff
Goes to the living room after lunch
Talking with her friends
Have her pills with her
No
Yes
Call SocialRobot (Voice command)
Recognized Kate's voice comand
Recognized Kate's location
Approaches Kate
Asks Kate "How can i help you?"
Responds "Get my pills"
Recognizes Kate's request
Goes to centre staff room
Requests Kate's pills
Gives Kate's pills to the SocialRobot
Receives the pills
Delivers the pills to Kate
Kate takes her pills

Figure 3.6: Flowchart of the Use Case Scenario 3

3.5 Discussion

The use case scenarios described in the previous sections introduce some of the functionalities intended for the SocialRobot, which enable it to accomplish a simple interaction with its users.

Many other possible scenarios can be addressed, regarding different roles for a social robot in many other situations. The SocialRobot project itself describes other possible scenarios [34], but they all converge to the goal of human-robot interaction.

Social robots can be integrated in peoples lives as companions for people in low social interaction environments, with the intention of giving them a social interaction and abstraction of loneliness situations, this scenario is mostly directed to the elderly population. Many scenarios in elderly care centers can also be exploited, in order to recognize abnormal behavior situations, provide information for services and assistance in daily tasks. In home scenarios the roles of a social robot can be perceived in the same perspectives, but the robot needs to have more autonomy, as it might be the only assistant for the elderly. In emergency situations the robot must be able to communicate with more capable entities.

Elderly population seams to be the greatest target for the services of a social robot, but regardless of the physical needs and illness constrains, every demographic class can use a social robot for commodity issues and house care scenarios. Looking to the present day, one of the most popular advances for computer interaction, is the Microsoft Kinect, which enables the recognition of gestures and movements in order to capture the user intentions that in the past were passed by an wired controller. This new approach has proven to provide a greater interaction and motivation for users to play a certain game. Since interaction also presumes communication, the integration of communication modules remains a constant requirement, as any robotic or computer based system should be able to understand human requests and commands, responding accordingly to the human needs. Home scenarios can also be approached as situations were humans just use the robot to perform simple tasks for them. Here the social awareness and autonomy of the robot is not the main concern and the system can even just be composed of a central processing unit, with no physical display, to provide the management of other devices.

Chapter 4

Implementation and Integration

4.1 Framework

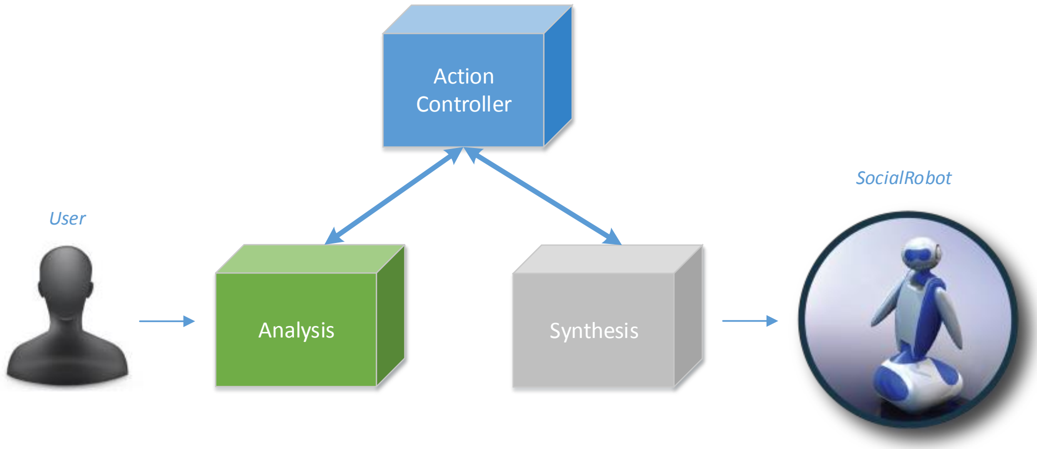


Figure 4.1: Framework diagram

The figure 4.1 represents the communication system overview. The communication system is concerned for the interaction between the SocialRobot and a user. The user is viewed as the input of the system, as he should be the focused part for the analysis. The SocialRobot is the interaction agent towards the user, equipped with synthesis modules, such as a speech synthesizer and an application launcher, in order to communicate and attend to the user's requests.

More detail about the analysis of the user interaction can be realized in figure 4.2. The microphones symbolize the input signals of the system, which should be disposed in the SocialRobot. By acquiring the audio signals from the microphone array, the system enters in the processing phase, where it will determine the user's location, voice commands and emotions for the analysis if the user's social behavior. Later on, the informations processed by the recognition modules are combined in order to choose between the speech recognition

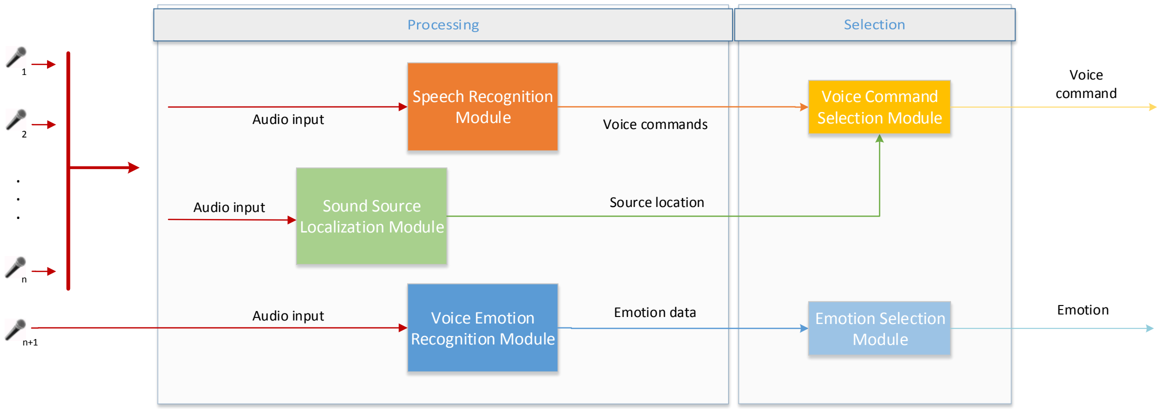


Figure 4.2: Analysis Diagram

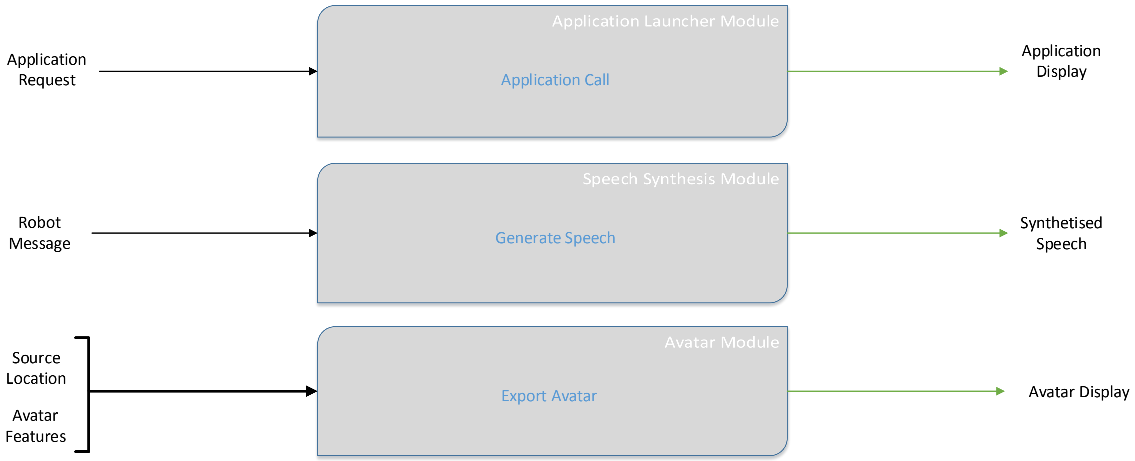


Figure 4.3: Synthesis Diagram

hypotheses and perform the social behavior analysis. The social behavior analysis is done inside the action controller, processing and crossing information between the user and the SocialRobot, in order to provide a natural way of communication between both parts and realize the corresponding actions to be performed by the synthesis system in figure 4.3.

4.2 Robot Operating System

4.2.1 Introduction

Originally developed in 2007 by the Stanford Artificial Intelligence Laboratory, with continued development at Willow Garage research institute till this year, ROS (Robot Operating System) provides libraries and tools to help software developers in the creation of robot applications. It provides hardware abstraction, device drivers, libraries, visualizers, message-passing, package management, and more. ROS is licensed under an open source, BSD license. ROS consists in a modular arrangement philosophy, for the organization of different programs and functionalities intended for a robotic system, which are executed in a peer to peer methodology. The programs for a robotic system, are considered as packages in the ROS architecture, and a set of packages are usually grouped into stacks. ROS packages include nodes, messages and services. A node represents the executable functionality of the package, and the messages consist of the information computed by the respective node. A robotic system can be perceived as a junction of nodes, which might have the necessity to communicate with each other. The communication between nodes is most commonly achieved with message passing through topics. Topics function as channels of communication between nodes that can be described as publishing and subscribing methods, e.g. when a node needs to send information to another node, the first one publishes a message containing that information to a topic and the second node subscribes that same message from the topic, as shown in figure 4.4.

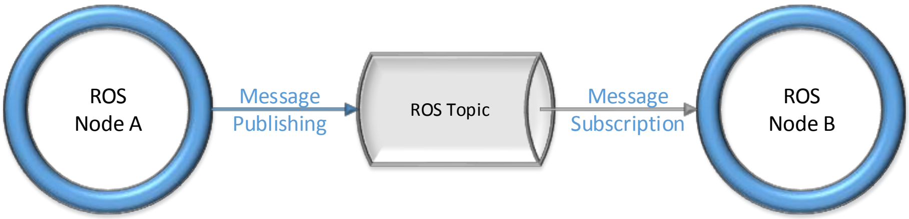


Figure 4.4: ROS nodes communication through a ROS topic.

Another feature of ROS is the support of multi-language programing, creating the ability for nodes to communicate even if they include different programing languages. ROS also comes with a base set of tools for system's management and visualization, categorized into command line tools and graphical user interfaces.

In terms of modularity, a system can be composed by various nodes, but some of those nodes can stay disconnected in order to use only some functionalities of the system. To accomplish this it is usual to use the roslaunch tool, which will be responsible for the nodes initialization, accordingly to a the user specifications contained in a launch file. Another useful tool is Rviz, that enables the user to visualize the contents of topics in a graphical way.

Gathering all its features, ROS reveals itself as a powerful tool for robotics research, with flexibility to integrate sensorial information and functionalities into any robotic platform.

4.2.2 ROS Integration Framework

The ROS framework developed for this thesis, described in figure 4.5, gives an overview of the integration in ROS of the packages inside the communication system. The image gives a better representation of the connections between the modules in the framework diagram and resumes the system's functionality.

The next sections describe the details of the packages introduced in the ROS integration diagram, their connections and message exchanges. All the packages can be perceived as ROS nodes, which share ROS topics in order to publish and subscribe their respective messages. Moreover, the structure of the messages is described in figure 4.6.

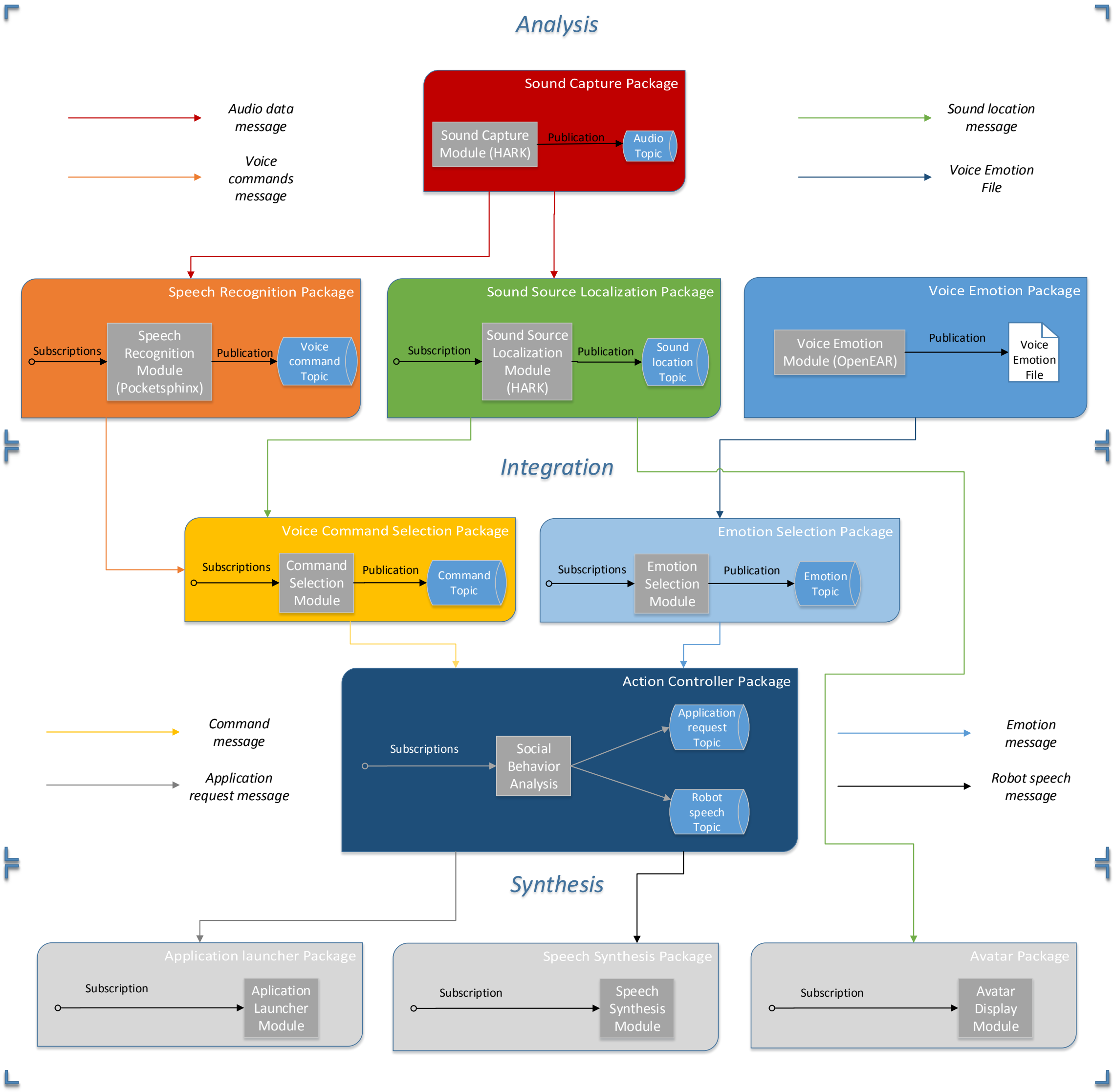
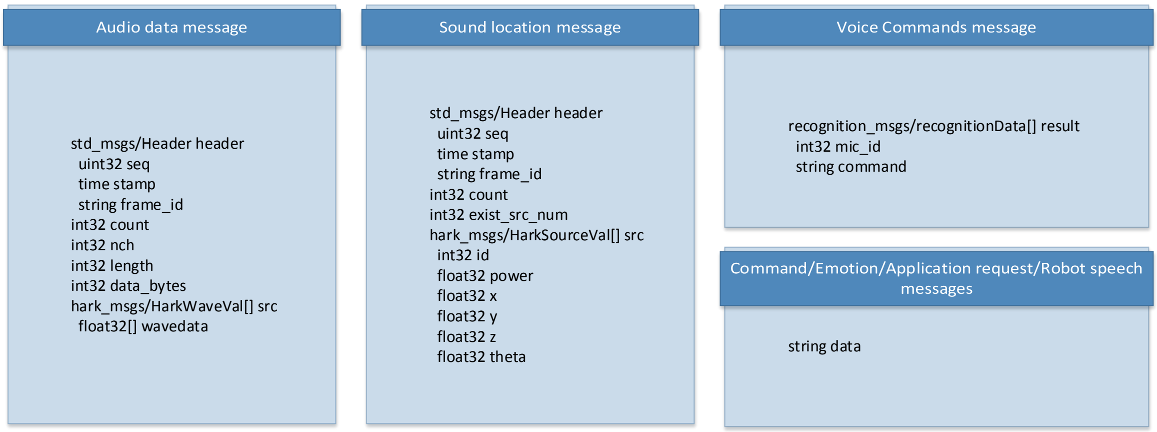


Figure 4.5: ROS integration diagram.



Audio data message

```
std_msgs/Header header
  uint32 seq
  time stamp
  string frame_id
int32 count
int32 nch
int32 length
int32 data_bytes
hark_msgs/HarkWaveVal[] src
  float32[] wavedata
```

Sound location message

```
std_msgs/Header header
  uint32 seq
  time stamp
  string frame_id
int32 count
int32 exist_src_num
hark_msgs/HarkSourceVal[] src
  int32 id
  float32 power
  float32 x
  float32 y
  float32 z
  float32 theta
```

Voice Commands message

```
recognition_msgs/recognitionData[] result
  int32 mic_id
  string command
```

Command/Emotion/Application request/Robot speech messages

```
string data
```

Figure 4.6: ROS message types and contents.

4.3 Sound Capture Package

The sound capture package is the most simple and also the most important package for the integration of the other packages. This package is intended for the Microsoft Kinect microphone array, but can easily be modified for any other hardware. There are many existing programs and applications to perform audio capture and record, but most of them are intended to create audio files. Thus a solution was needed to capture a continuous audio stream. Another challenge was the detection of the Microsoft Kinect as an audio device in Ubuntu. The HARK toolkit solved those both problems by delivering a Kinect audio driver and tools that enabled the capture of continuous multichannel audio streams, which can be sent to all the other packages that will need them. To capture sound, an HARK network was created, see figure 4.7, using HARK's FlowDesigner tool, and configured accordingly to the Kinect's microphone array specifications. The network can be considered as an executable file, responsible for the creation of the audio node and the transmission of the audio data to the respective ROS topic.

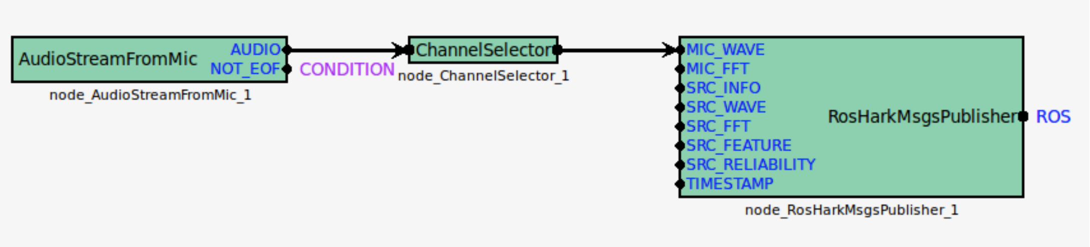


Figure 4.7: Sound Capture Network

This approach worked, but it was not the ideal solution. The integration with other packages, brought the necessity to inclose the HARK network into a ROS package of its own, in order to enable the use of ROS based tools.

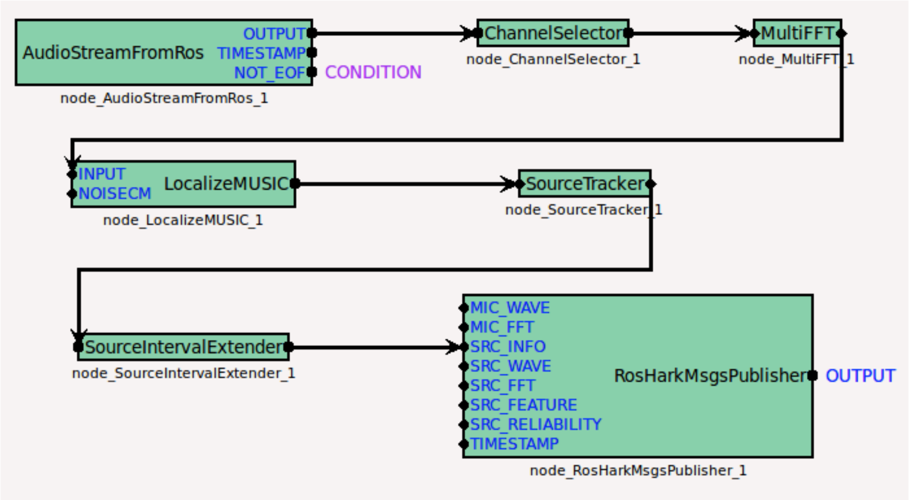


Figure 4.8: Sound Source Localization Network

4.4 Sound Source Localization Package

In order to create a recipe to estimate the location of a sound source, sound signals are the main ingredient. As noted previously, there is a package responsible for sound acquisition, sending the captured audio data into a ROS topic. The sound source localization (SSL) package acquires the audio data by subscribing it from that same topic. This package is also integrated using the HARK toolkit, which makes its development a similar process, compared to the sound capture package. Once again an HARK network was created , but this time we had to add some extra capabilities, see figure 4.8.

The most important module for the localization process is the LocalizeMUSIC, which requires a transfer function file containing the positional relationship between the microphones and possible sound sources. This module is responsible for the detection of the direction and power of a sound source. This configuration used for this thesis only considers one sound source, but this package can bring the ability to detect up to four simultaneous sound sources, which can be used for other scenarios, where a source separation must be performed.

The sound source localization package outputs the direction of arrival (DOA), the cartesian coordinates and the spectrum power of the sound source.

All the informations about the sound source location are joint in a ROS message, which is published to a topic for posterior access.

4.4.1 HARK

HARK is an audio processing toolkit capable of capturing continuous multichannel audio streams from microphone arrays and performing sound source localization. In order to perform sound source localization, HARK delivers the LocalizeMUSIC module along with the harktool4, for the creation of transfer functions files consisting in a steering vector for the positional relationship between microphone arrays and possible sound sources, see figure 4.9.

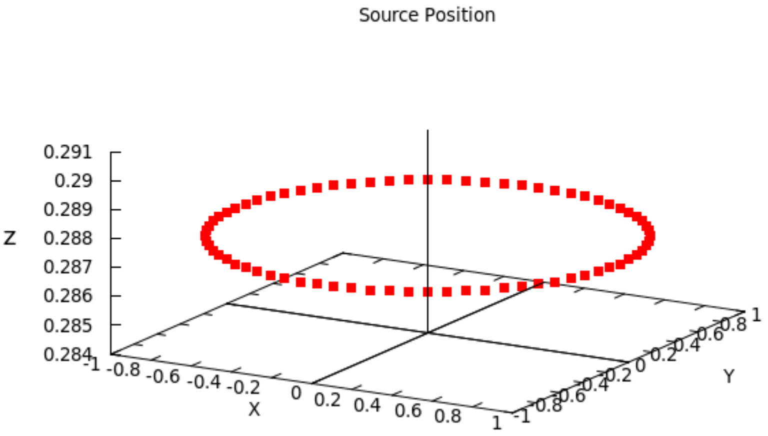


Figure 4.9: Graphical representation of a transfer function file, containing 72 possible sound sources for the SSL task, created with the harktool4.

The transfer function from sound to each microphone can be measured or calculated by specifying the coordinates of the microphones and the the amount of sound sources desired for the SSL task within a direction range in the horizontal plane, with the harktool4. The transfer function file is used by the LocalizeMUSIC to include an a priori information of the recognizable sound sources. The LocalizeMUSIC module also requires a complex frequency representation of the audio input signals. For that purpose, HARK has the MultiFFT module that converts the multichannel waveform data into spectra and analyses the spectra in the time frequency domains.

The LocalizeMUSIC computation is based in the MUSIC method, which estimates the power and the DOA of a sound source using algorithms for eigenvalue decomposition of the correlation matrix (CM) among input signal channels. Once sound source localization package is intended to work in real environments, we used the HARK's SourceTracker module to define a power threshold that must be exceeded on order to activate the SSL task. The SourceTracker also brings the ability to track sound sources for as long as their power stays above the defined threshold.

For the sound source localization package, we used the Standard Eigen Value Decomposition (SEVD) MUSIC algorithm and assumed the existence of only one sound source for every SSL iteration.

Parameter Name	Type	Used Value	Unit	Description
MUSIC_ALGORITHM	string	SEVD		Algorithm of MUSIC
TF_CHANNEL_SELECTION	Vector<int>	<Vector<int> 0 1 2 3>		Used channels
LENGTH	int	512	pt	FFT points (NFFT)
SAMPLING_RATE	int	16000	Hz	Sampling rate
A_MATRIX	string	"kinect_loc.dat"		Transfer function file name
WINDOW	int	50	frame	Frames to normalize CM
WINDOW_TYPE	string	FUTURE		Frame selection to normalize CM
PERIOD	int	50	frame	Cycle to compute SSL
NUM_SOURCE	int	1		Number of sound sources
MIN_DEG	int	-180	deg	Minimum azimuth
MAX_DEG	int	180	deg	Maximum azimuth
LOWER_BOUND_FREQUENCY	int	500	Hz	Lower bound frequency
UPPER_BOUND_FREQUENCY	int	2800	Hz	Upper bound frequency
SPECTRUM_WEIGHT_TYPE	string	Uniform		Type of frequency weight
ENABLE_EIGENVALUE_WEIGHT	bool	false		Enable eigen value weight
DEBUG	bool	false		ON/OFF of debug output

Table 4.1: Parameter list for the LocalizeMUSIC module

A simple description of the LocalizeMUSIC parameters is given in table 4.1, also presenting its configuration values for the sound source localization package. All the configurable parameters of HARK's LocalizeMUSIC module are detailed in [21], along with a mathematical description for the MUSIC method and all the algorithms presented for this toolkit.

4.5 Speech Recognition Packages

For the speech recognition system implementation, two approaches have been considered, resulting in two different ROS packages.

The packages differ on the sound capture method for the speech recognizer. The first method consist on the sound capture of a single audio stream from all the microphones that compose the Kinect's microphone array. The second method also incorporates all audio streams from the microphone array, but they are analyzed separately. This method takes advantage of the sound capture package, subscribing the audio data from each microphone.

The two approaches for the speech recognition packages integrate the CMU-Sphinx toolkit, using the Pocketsphinx recognizer. Both packages require an acoustic model, language model and a dictionary file, in order to initialize the Pocketsphinx decoder.

The first package, which captures the combined audio stream, has no dependency with the sound capture package, acquiring the audio stream directly from the microphone array. The recognition process is triggered by speech detection, enabling the recognizer to compute utterances with no length limit. This technique is usually used for dictation systems or recognition of long speech sentences. This speech recognition package, only gives one recognition result, due to its single audio stream input. The result is inserted in a ROS message and then published into a ROS topic, making it available for subscription from other packages. This package is available in ROS [4], but it needed some modifications in order to capture sound with the Kinect and adjust to the intended voice commands to be recognized.

The goal for the integration of the speech recognition system in this thesis focus on the recognition of voice commands instead of dictation purposes. The implemented recognition process analyses four audio signals separately, in order to increase the system's accuracy by relating the results with the recognized angles given by the sound source localization package. Although this system purpose relies in the recognition of voice commands, online considerations are also required, such as voice detection algorithms for audio segmentation, in order to perform continuous speech recognition. Furthermore, this second package joins a voice detection algorithm with the capacity to analyze each channel of the microphone array, which improves the recognition of the voice commands for non-frontal positions. This package outputs four recognition results, each from each microphone. The results are also inserted in a ROS message, containing not only the four recognition results but also the microphone identifier for each obtained result.

4.5.1 CMU-Sphinx

The CMU-Sphinx is a HMM-based speech recognition system, requiring training and decoding processes. The training process is performed by learning the characteristics of a set of sound units. The decoding process is base on the knowledge acquired from the training process, which will be used to find the most probable sequence of sound units for a given speech signal. The CMU-Sphinx provides tools for both processes, such as SphinxTrain and Pocketsphinx. The training tools learn the parameters of the sound unit models, using a set of sample speech signals. It also needs to know which sound units it needs to learn and the sequence in which they occur in every speech sample. This information is given through a transcript file, in which the sequence of words and non-speech sounds are written exactly as they occurred in a speech sample, followed by a tag which can be used to associate the sequence with the corresponding speech signal. The training tools also needs a language dictionary and a filler dictionary, regarding the words mapped with sequences of sound units and non-speech sounds mapped with the corresponding sound units.

Fortunately, due to development state of speech recognition systems, and CMU-Sphinx itself, many pre-trained acoustic models are already incorporated into this toolkit, abstracting researchers from the need of data acquisition and posterior training.

For this thesis the selected acoustic model was de WSJ [43], trained through the Wall Street Journal training corpus. This acoustic models introduces a set of 40 Hidden Markov

Model (HMM) states, 39 monophones models and 1 silence model. Each HMM has three output states with a left-to-right topology with self-loops and no transitions which skip over states.

Speech recognition systems perform a maximum a posteriori estimation, to retrieve the the most likely sequence of words from a speech utterance, given the sequence of feature vectors extracted from the speech signal. Those feature vectors for the classification process consist of the Mel-frequency cepstral coefficients (MFCCs) [5].

$$Word_1 \cdots Word_n = argmax_{Wd_1 \cdots Wd_n}\{P(feature\,vectors|Wd_1 \cdots Word_n)P(Wd_1 \cdots Wd_n)\} \tag{4.1}$$

where $[Word_1 \cdots Word_n]$ is the recognized sequence of words and $[Wd_1\,Wd_n]$ is any sequence of words. The argument of the right side of the equation 4.1 is composed by the probability of feature vectors, given a sequence of words $P(feature\,vectors|Wd_1 \cdots Wd_n)$, provided by the HMMs, and the probability of the sequence of words $P(Wd_1 \cdots Wd_n)$, provided by the language model.

As it was chosen a pre-trained acoustic model, the HMM learning was not concerned in this thesis, and with the Pocketsphinx decoder, the HMM classifier is also taken care of. However, HMMs can be perceived as a probabilistic method for the classification process, which identifies a word based on the transition of states. Each state corresponds to a word's sound unit, and each word has its characteristic sequence of sound units, thus the recognition result of a sequence of sound units will match trained word in the dictionary file, and that word will be the best recognized word hypothesis with a certain probability. For a more detailed information about HMMs, see [35].

N-gram language models, assume that the probability of any word in a sequence of words depends only on the previous N words in the sequence. Thus the calculation for $P(Wd_1 \cdots Wd_n)$ for a N-gram model:

$$P(Wd_1 ... Wd_n) = P(Wd_1)P(Wd_2|Wd_1)P(Wd_3|Wd_2, Wd_1) \cdots P(Wd_n|Wd_{n-1}, \cdots, Wd_1) \tag{4.2}$$

The speech recognition system for this thesis, is only intended to recognize voice commands, abstracting dictation purposes. Thus, the language model can be restricted to 3-gram models:

$$P(Wd_1 ... Wd_n) = P(Wd_1)P(Wd_2|Wd_1)P(Wd_3|Wd_2, Wd_1) \cdots P(Wd_n|Wd_{n-1}, Wd_{n-2}) \tag{4.3}$$

The language model created for this thesis speech recognition system, is based on a speech corpus file containing the recognizable voice commands. The corpus file is used as input for CMU lmtool, which creates the 3-gram language model. This tool also creates a dictionary file, where words are mapped into their corresponding phoneme sequences.

4.6 Voice Emotion Package

Although the openEAR toolkit can be considered as an out-of-the-box emotion recognition software, in order to accomplish the user's social behavior analysis, the recognized emotions need to be accessible for a combined analysis with the other packages output. Thus, the emotion recognition module, besides the openEAR recognition engine, also needs its own ROS package to send the emotion messages to the action controller package.

Due to openEAR's complexity and dependences, its integration in ROS was achieved by altering the openEAR's source code and configuration files, in order to export the recognized emotions to a file that could be accessed from another ROS package. Nevertheless, arrangements were made to maintain the live recognition functionality. The openEAR's authors designed the live recognition option to use the PortAudio library, thus it maintains a permanent access to the audio device, blocking the sound capture package and the audio data messages needed for the other analysis packages. Furthermore, to pass this inconvenient, another device for the sound capture was added. This solution was also found in a similar work [24].

For the emotion recognition process we used the openEAR's continuous dimensional labels for arousal and valence in the range from -1 to +1, trained from the Sensitive Artificial Listener (SAL) corpus [13], which are exported into an emotion data file for a posterior clustering into emotional quadrants, realizing active-positive, passive-positive, active-negative and passive-negative emotional states. However, this package can execute another ROS node, capable of exporting the results for other emotion models, considering basic emotion classes.

The new audio device can also be a multichannel recording device, but this toolkit uses its own approach to optimize the recognition results. When a multichannel device is chosen as the audio recording device, all the audio channels are mixed down into a single channel, which will be analyzed in the feature extraction tool, in order to realize the best features for the recognition process.

4.6.1 OpenEar

OpenEAR is an affect and emotion recognition toolkit for audio and speech affect recognition. The openEAR toolkit core component is the SMILE (Speech and Music Interpretation by Large-Space Extraction) signal processing and feature extraction tool. OpenEAR can be used as an out-of-the-box emotion live affect recognizer for various domains, using pre-trained models which are included in the distribution [14]. The most interesting model-sets provided with this toolkit, were trained on emotion speech databases, such as the Berlin Speech Emotion Database (EMO-DB) [10], the Airplane Behavior Corpus (ABC) and the Audio Visual Interest Corpus (AVIC) [38]. The EMO-DB based model includes seven classes of basic emotions: anger, fear, happiness, disgust, boredom, sadness and neutral. The ABC emotion classes are labeled as aggressive, cheerful, intoxicated, nervous, neutral and tired. Finally, the model trained from the AVIC database, concerns about the speaker interest levels, categorized in disinterest, normal and high interest classes. This toolkit also addresses continuous dimensional labels for valence and activation in the range from -1 to +1, trained from the Sensitive Artificial Listener (SAL) corpus and the Vera am Mittag (VAM) corpus [17]. To achieve emotion recognition, the signal input can either be read offline from audio files or recorded online from a audio capture device. The sound capture from audio devices is performed using the PortAudio library for live audio input, enabling real-time audio features extraction and emotion classification. The SMILE feature extraction tool is capable of extracting low-level audio features (Low-Level Descriptors (LLD)) and applying various statistical functionals and transformations to those features. The Low-Level Descriptors are listed in table 4.2.

Feature Group	**Description**
Signal Energy	Root mean-square & logarithmic
FFT-Spectrum	Bins $0 - N_{fft}$
Mel-Spectrum	Bins $0 - N_{mel}$
Cepstral	MFCC $0 - N_{mfcc}$
Pitch	Fundamental frequency F_0 via ACF, in Hz, bark and closest semitone Probability of voicing $\frac{ACF(T_0)}{ACF(0)}$
Voice Quality	Harmonics to noise ratio
LPC	Linear Predictive Coefficients
PLP	Perceptual Linear Predictive Coefficients
Formants	Formants and Bandwidth computed from LPC analysis
Time Signal	Zero-crossing rate, maximum value, DC
Spectral	Energy in bands (Hz), N% roll-off point, centroid, flux and relative position of spectrum max and min.
Musical	CHROMA (warped semitone filter bank), CENS (Comb-filter bank)

Table 4.2: Low-Level Descriptors in openEAR's SMILE feature extractor.

Live demonstrators for audio processing tasks often require segmentation of the audio stream, thus this toolkit also provides voice detection algorithms and a turn detector. For incrementally classifying the features extracted from the segments, Support Vector Machines are implemented using the LibSVM library [11].

4.7 Synthesis Packages

The communication system intends to establish an interaction between the SocialRobot and its users, thus it could not be designed only to comprehend the user's demands and social behavior. Furthermore, it also needs to enable the SocialRobot to express itself towards its users and accomplish the execution of some tasks.

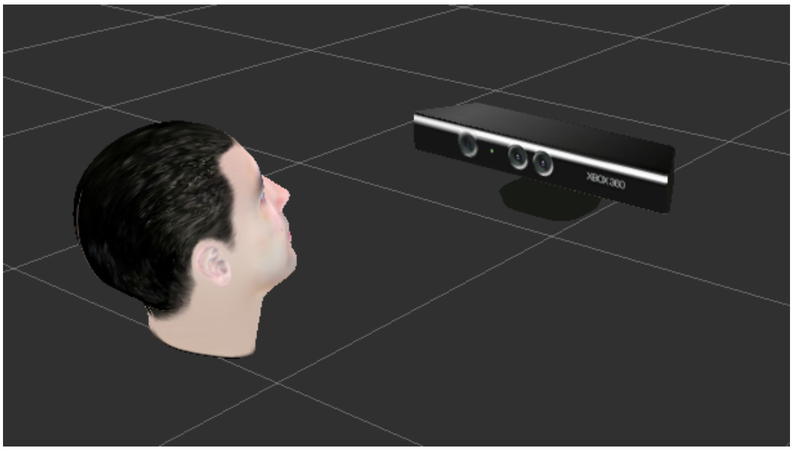


Figure 4.10: Sound Source Localization results display.

According to the use case scenarios, the tasks realized for the system intend the execution of applications that allow interaction with other people and health care services, such as Skype[1]. Thus we created the **Application Launcher Package** to execute such applications. The package functionality can be resumed from its ability to receive application requests and their accomplishment. From a more detailed perspective, this package subscribes an application request message from a ROS topic, in order to realize which application to launch.

Considering that the robot also needs the ability to communicate with the user, we also developed a package to perform speech synthesis. The **Speech Synthesizer Package** also subscribes a message from a ROS topic, containing the proper answers and questions for the user, exporting them as synthesized speech.

The last synthesis package is the **Avatar**, which consists in a 3D visualization of an human face model that assumes the direction of the recognized sound source from the user, relatively to the Kinect's microphone array. This package only serves visualization purposes, but it could also be perceived as an interface feature, as shown in figure 4.10.

4.8 System Integration

After the development of the above ROS packages and the integration of the recognition toolkits used in the user analysis packages, it was also necessary to integrate their outputs. All the packages have modular characteristics for their own purposes, but in order to create a social aware communication system they are related to each other, due their information exchange, as shown in figures 4.5 and 4.6.

We start by capturing the audio signals, representing the user's speech utterances, and publishing them as audio data messages into the audio topic, which will be subscribed by the sound source localization and speech recognition packages, in order to initiate their

[1] Free application for internet calls and messaging - http://www.skype.com

recognition processes. At the same time, the voice emotion package is also acquiring the same speech utterances, directly from another audio recording device, writing the recognized values of arousal and valence into a emotion data file, which will be analyzed by the emotion selection package, in order to realize the corresponding emotional state, as described in subsection 4.8.2.

Since all of the recognition packages are equipped with voice detection algorithms, the system's live functionality is preserved, presenting one recognition result from each package for each detected utterance. However, our approach for the speech recognition package introduces four recognition results for the same utterance, from which the best recognition result is resolved by the voice command selection package, detailed in subsection 4.8.1.

The social behavior analysis is mostly comprehended through the speech and emotion recognition results. However, to accomplish a communication system capable of providing an interaction between the user and the SocialRobot, we needed to introduce the action controller package to analyze those results and manage the action capabilities of the synthesis packages, as described in subsection 4.8.3.

4.8.1 Voice Command Selection Package

Once we have four speech recognition outputs, one for each Kinect's microphone, our choice between the four hypotheses comes from assuming that each microphone as an optimal direction range for the recognition process. Thus we integrate the speech recognition results with the sound source localization results. That integration is performed by the voice command selection package, which subscribes those results and publishes the best speech recognition hypothesis, according to a predefined range of directions that determine the best audio signal captured from the Kinect's microphone array, as realized in figure 4.11.

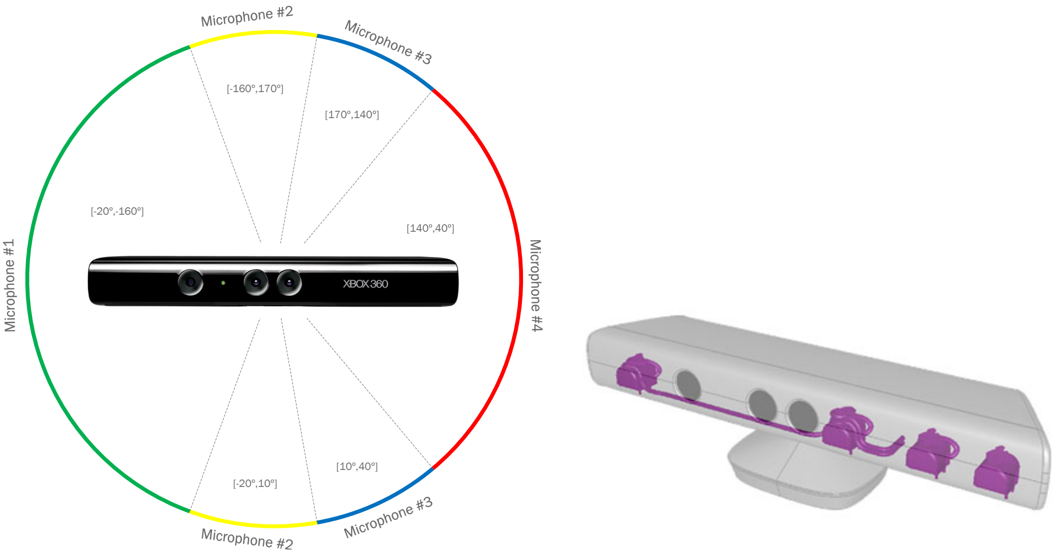


Figure 4.11: Left side - Best angle range for each Kinect's microphone. Right side - Kinect microphone array disposal.

4.8.2 Emotion Selection Package

Emotion recognition applications usually label the user's emotional state into rough classes of basic emotions, such as happy, sad, fear, anger and surprise. However, those classes often do not optimally reflect emotions that typically occur in a given scenario [46]. Since the use case scenarios intend the recognition of distress signals, which cannot be optimally clustered by the usual basic emotion classes, we decided to analyze the user's emotion by mapping the arousal and valence values, from the emotion data file, into quadrants, as in figure 4.12. Those quadrants realize emotional states as positive-active, positive-passive, negative-active and negative-passive, following a similar approach to the one described by Wöllmer et al. [46]. Furthermore, a functional ROS package was accomplished with message publishing capabilities, which makes the emotion recognition results available in a ROS topic that can be subscribed from other packages to perform the social behavior analysis.

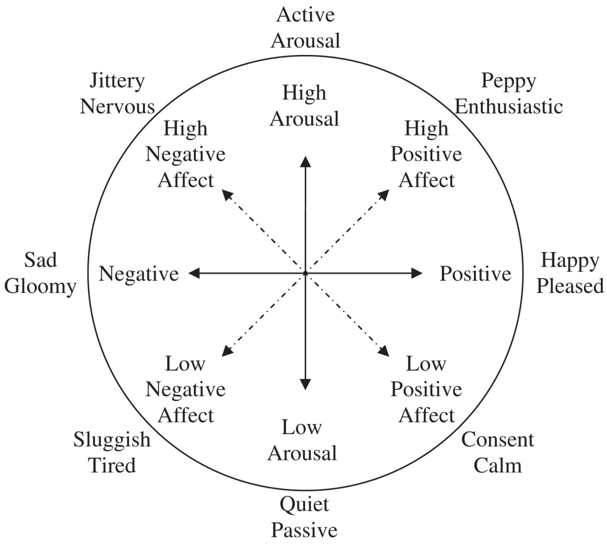


Figure 4.12: Graphical representation of the affective circumflex, with the valence dimension represented horizontally on the x-axis and arousal vertically on the y-axis [3].

4.8.3 Action Controller Package

In order to attain the challenges presented by the use case scenarios, communication protocols between the user analysis and the SocialRobot had to be considered, as introduced in figure 4.1. Thus, when the robot is presented with a task given by the user, it should give a valid answer or question for more information about the user's intentions. If the robot asks a question to the user, it should also wait for a matching answer.

The use case scenarios describe situations were the robot is communicating with a person, passing through states of knowledge until performing a final interaction act. Moreover, the system needs to realize the users social behavior and recognize their demands, performing the respective actions.

The action controller joins the informations given by the voice command and emotion selection packages, evaluating the user's social behavior and exporting the respective actions to the synthesis packages. Those actions set the start of the speech synthesizer and the application launcher, to give answers and ask questions to the user and launch applications, when such requests are perceived. Since the action controller was designed to comprehend the hypothetical situations described in the use case scenarios, for demonstration purposes, the interaction stages are realized when the respective emotional states and voice commands are recognized, i.e, the action controller realizes pre-defined actions from combinations of the speech and emotion recognition results, as described in tables 4.3, 4.4 and 4.5.

From a ROS perspective, the action controller subscribes the command and emotion messages, published by the voice command and emotion selection packages, publishing application requests and robot speech messages into their respective topics, that later on will be executed by the synthesis packages.

Interaction #	Speech Recognition	Emotion Recognition	Action
1	Help	Active-Negative	Ask the user if he wants to request Help
2	Yes	Passive-Negative	Launch Skype

Table 4.3: Use Case Scenario 1 Interaction Stages

Interaction #	Speech Recognition	Emotion Recognition	Action
1	-	-	Ask the user if he is feeling well
2	No	Passive-Negative	Ask if he has stomach pain
3	Yes	Passive-Negative	Launch Skype

Table 4.4: Use Case Scenario 2 Interaction Stages

Interaction #	Speech Recognition	Emotion Recognition	Action
1	Come here	-	Ask the user what the SocialRobot can do to help
2	Get my pills	-	Request the caregiver or center staff

Table 4.5: Use Case Scenario 3 Interaction Stages

Chapter 5

System Results

This chapter includes all the tests and evaluations performed to determine the system's performance. Since the communication system relies on three major packages, all responsible for the user analysis, we decided to evaluate each package's performance individually. In the individual evaluation for each package, more exhausting tests were conducted to obtain more deterministic results.

The integration results are presented in section 5.2, demonstrating the accomplishment of the use case scenarios.

5.1 Individual Results

- **Sound Source Localization**

The experimental setup for the sound source localization tests consisted in the live recognition of the direction of arrival (DOA) of sound sources, generated by a human speaker from different directions. In order to minimize possible errors, the speaker was placed on marked spots according to the measured angles in a one meter radius. Since the microphone array is located in the front of the Kinect, when the sound comes from behind the device, it suffers some deflections that have an affect on the recognition process, thus we limited the angle range for the recognition trials from -120^{o} to 120^{o} degrees.

From the results presented in table 5.1, we can realize the system's precision in the localization of the speech directions. The results show a good precision for the recognition process, but also a fixed error of five degrees. This error is related to the transfer function generation process, where 72 sound sources are considered from a 360^{o} range and separated by five degrees in the horizontal plane. Thus, the recognizable sound source directions differ exactly five degrees from each other, imposing a maximum precision of five degrees. Since the localization system is not intended for high precision measures, we consider it as a reasonable standard deviation. Nevertheless, we could consider more sound sources in the transfer function for higher precision rates.

Measured Angle	Recognized Angles (30 Trials)	Average Recognition
−120º	7 × −115º 23 × −120º	118.33º
−90º	27 × −90º 3 × −95º	90.5º
−45º	29 × −45º 1 × −50º	45.17º
0º	30 × 0º	0º
45º	2 × 40º 28 × 45º	44.67º
90º	4 × 85º 26 × 90º	89.33º
120º	9 × 115º 11 × 120º 10 × 125º	120, 17º

Table 5.1: Results for Sound Source Localization.

Figure 5.1 presents a graphical representation of the system's accuracy, from which we can realize that the frontal positions are the ones with the higher exact angle recognition rates. It is also perceivable a decrease of the recognition rates for the positive sound sources, due to the microphone array disposal, from which the audio signals will suffer more deflections between the last pair of microphones, resulting in the emphasis of the five degrees error. However, if we consider the recognitions with a maximum deviation of five degrees as successful recognition results, the system accomplishes a perfect recognition rate for all the tested directions.

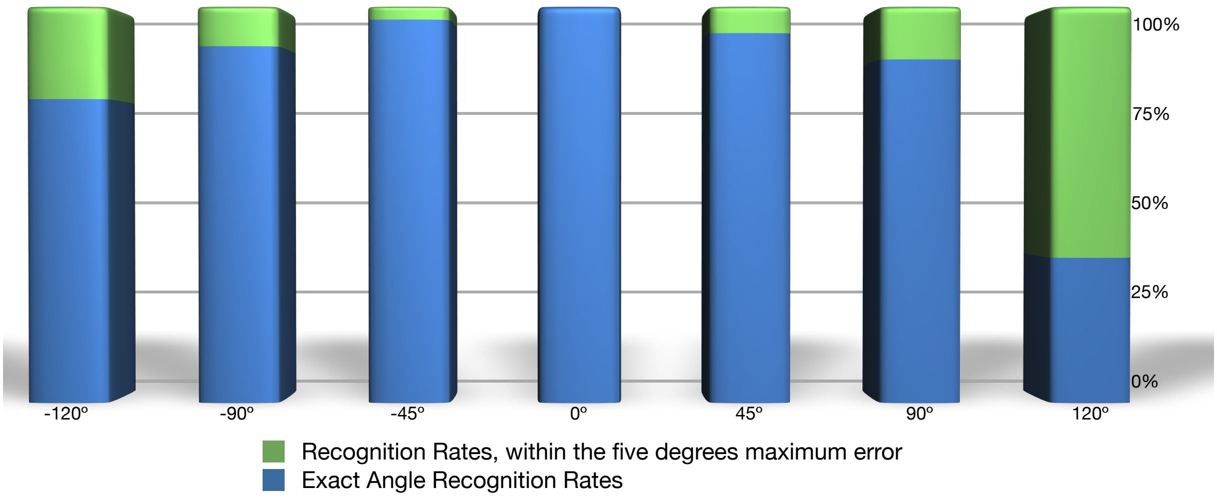


Figure 5.1: Sound Source Localization results.

- **Speech Recognition**

The speech recognition and sound source localization results are integrated, in order to test our speech recognition approach. The experimental setup is similar to the one used for the sound source localization tests, but since this system needs to be suited for various users and uses the WSJ acoustic model, we decided to use high quality synthesized American voices.

To test the speech recognition system, we created a dataset containing five female and five male synthesized American voices for four different directions, simulating the recognizable voice commands considered for the dictionary file and language model. Since our sound capture package is capable to record wave files for each Kinect's microphone, we manually chose the best wave file associated to the respective microphone for each trial, according to the optimal range defined for each microphone, as shown in figure 4.11. Table 5.2 presents the results for those tests, from which we can realize that our approach lead to good recognition results for all the directions tested, removing the spatial dependency of the speech recognition system.

	Angles and corresponding recognition rates			
Voice Commands	−60º **VCRR(%)**	−10º **VCRR(%)**	30º **VCRR(%)**	70º **VCRR(%)**
Yes	90%	100%	100%	100%
No	90%	100%	90%	90%
Hello	70%	70%	80%	70%
Help	50%	20%	60%	70%
Goodbye	90%	90%	90%	100%
Come here	100%	100%	100%	90%
Launch Skype	80%	100%	100%	100%
Get my pills	70%	90%	80%	100%

Table 5.2: Voice command recognition for speech signals from specific angles. VCRR (Voice Command Recognition Rate)

From the analysis of the figure 5.2, we can perceive good recognition rates for the tested voice commands, apart from the "Help" and "Hello" commands. Moreover, the recognition rates for the "Help" command are the lowest, but the most wrongly recognized word for this command was "Hello". Those wrong recognitions can be justified by the words similarity, which brings the need to enhance the captured speech signals for better results in such cases. The enhancement of the speech signals can be achieved from noise reduction methods and better voice detection algorithms.

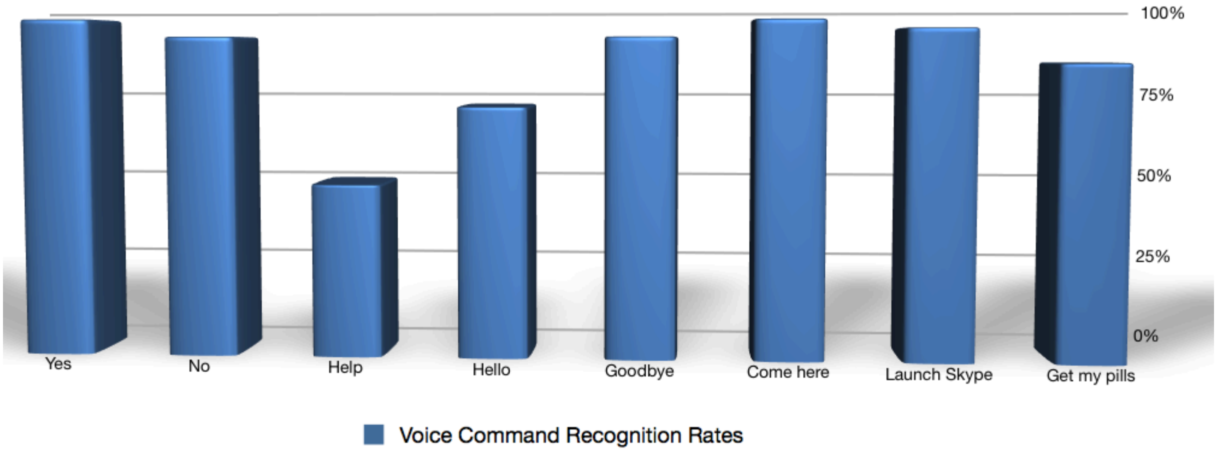


Figure 5.2: Speech recognition overall recognition rates.

- **Voice Emotion**

Testing a voice emotion recognition system is challenging, since we have to simulate emotional speech. Usually, the not so good quality of the simulated emotions leads to bad results, particularly if they don't apply to the trained model standards. The most used approach to surpass this issue is to use of emotional speech databases. One of the most used databases is the Berlin Database of Emotional Speech (EMO-DB) [10] and since the openEAR has an emotion model trained from this database, we decided to use it to evaluate the openEAR's stand-alone performance, and applicability to more specific applications. The results from table 5.3 show some lower recognition rates for the neutral and fear classes. However, since the communication system analyzes the user's interaction beyond the emotion recognition results, considering them a step towards the perception of the user's behavior, the action controller package can be designed to comprehend false-positive emotion recognitions and maintain a reliable interaction with the SocialRobot's users. Nevertheless, we can overcome this issue by improving the emotion recognition models or even include a different emotion recognition system, once the communication system does not rely on any specific software but only on their outputs, due to its modular architecture.

		Recognized Emotions						
		Happiness	**Sadness**	**Neutral**	**Fear**	**Anger**	**Disgust**	**Boredom**
Actual Emotions	**Happiness**	76.06 %	0.00%	0.00%	1.41%	21.12%	1.41%	0.00%
	Sadness	0.00%	93.55%	0.00%	0.00%	0.00%	1.61%	4.84%
	Neutral	6.33%	2.53%	65.82%	3.80%	0.00%	0.00%	21.52%
	Fear	15.94%	5.80%	1.45%	65.22%	10.14%	1.45%	0.00%
	Anger	3.15%	0.00%	0.79%	0.00%	95.28%	0.79%	0.00%
	Disgust	0.00%	2.17%	2.17%	2.17%	2.17%	89.13%	2.17%
	Boredom	0.00%	8.64%	3.70%	0.00%	0.00%	1.23%	86.42%

Table 5.3: Confusion table of the emotion recognition results for the EMO-DB model, including seven emotion classes.

We used the models containing continuous dimensional labels for valence and arousal, clustering emotions as positive-active, positive-passive, negative-active and negative-passive by mapping them into quadrants, review figure 4.12, we also had to test the results of our approach. Table 5.4 represents those results that were achieved by online testing the emotion recognition package with the voice commands considered for the use case scenarios, performing 25 repetitions for each emotional state.

		Recognized Emotional State			
		Active-Positive	**Active-Negative**	**Passive-Positive**	**Passive-Negative**
Actual Emotional State	**Active-Positive**	84 %	4%	12%	0%
	Active-Negative	32%	60%	4%	4%
	Passive-Positive	4%	0%	64%	32%
	Passive-Negative	0%	0%	24%	76%

Table 5.4: Confusion table of the emotion recognition results for the arousal-valence approach.

The results for our approach achieve an overall recognition rate of 71%, which is close to the 71.8% rate presented for the similar approach in [46]. However, we have to notice that their results consider tests using the emotional speech databases, while we tested our approach with live emotional voice commands. Furthermore, the short voice commands hardly realize emotional turns, which also influence our results.

From our results we can also perceive lower recognition rates for the active-negative and passive-positive emotional states. Those lower rates come from the arousal's and valence's relation with the signal's amplitude, from which a louder emotional speech is often associated with positive valence and a quiet emotional speech with negative valence. However, to overcome this issue, we could introduce longer emotional utterances, presenting more discriminative data and emotional turns.

5.2 Integration Results

In this section, the three use case scenarios were tested, according to their interaction stages described in chapter 3.

Notice that this results comprehend the hypothetical situations described for each of the use case scenarios, thus the interaction stages are performed when the respective emotional states and voice commands are recognized.

Since the interaction depends on the combination of the recognized voice commands with the associated emotional states, the success rates for the use case scenarios are directly related with the individual results of the speech and emotion recognition systems.

The success rates for the accomplishment of the use case scenarios, in table 5.5, evaluate the system's performance over twenty simulations for each use case scenario, concluding a better than chance probability to achieve the expected interaction with the communication system. Nevertheless, if we consider few repetitions for the voice commands with lower recognition rates, the use case scenarios success rates can improve significantly, e.g, in the first use case scenario simulations, the "Help" command with and active-negative emotional state is not always recognized at the first try. Furthermore, improving the speech and emotion recognition rates will also improve the success rates for the use case scenarios.

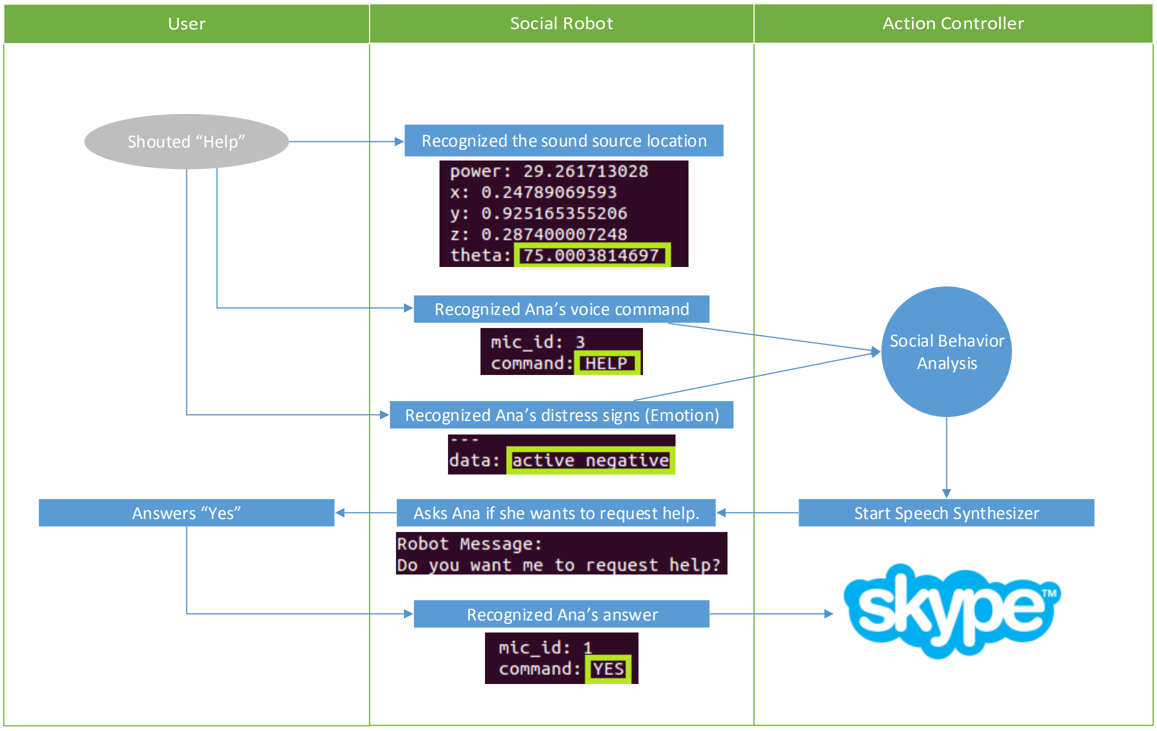
Each simulation consisted in a live interaction of a user with the communication system. That interaction was conducted according to the use case scenarios, with one user expressing voice commands from different directions with different emotional states, as considered for the hypothetical situations for each use case scenario.

	Success Rate
Use Case Scenario 1	60%
Use Case Scenario 2	75%
Use Case Scenario 3	80%

Table 5.5: Success rates for the use case scenarios simulations.

- **Use Case Scenario 1**

Recalling figure 3.2, for this use case scenario the SocialRobot is intend to recognize the "Help" command spoken by the user, with a negative emotional sate, and the sound source. Later on, the SocialRobot asks the user if she wants to request help and given an affirmative answer it performs a video call and alerts the user's neighbors. Through the simulation represented in figure 5.3, it is possible to realize the accomplishment of the use case scenario, resulting in the Skype initialization.


User
Social Robot
Action Controller
Shouted "Help"
Recognized the sound source location
Recognized Ana's voice command
Recognized Ana's distress signs (Emotion)
Social Behavior Analysis
Answers "Yes"
Asks Ana if she wants to request help.
Start Speech Synthesizer
Recognized Ana's answer
skype

```
power: 29.261713028
x: 0.24789069593
y: 0.925165355206
z: 0.287400007248
theta: 75.0003814697
```

```
mic_id: 3
command: HELP
```

```
---
data: active negative
```

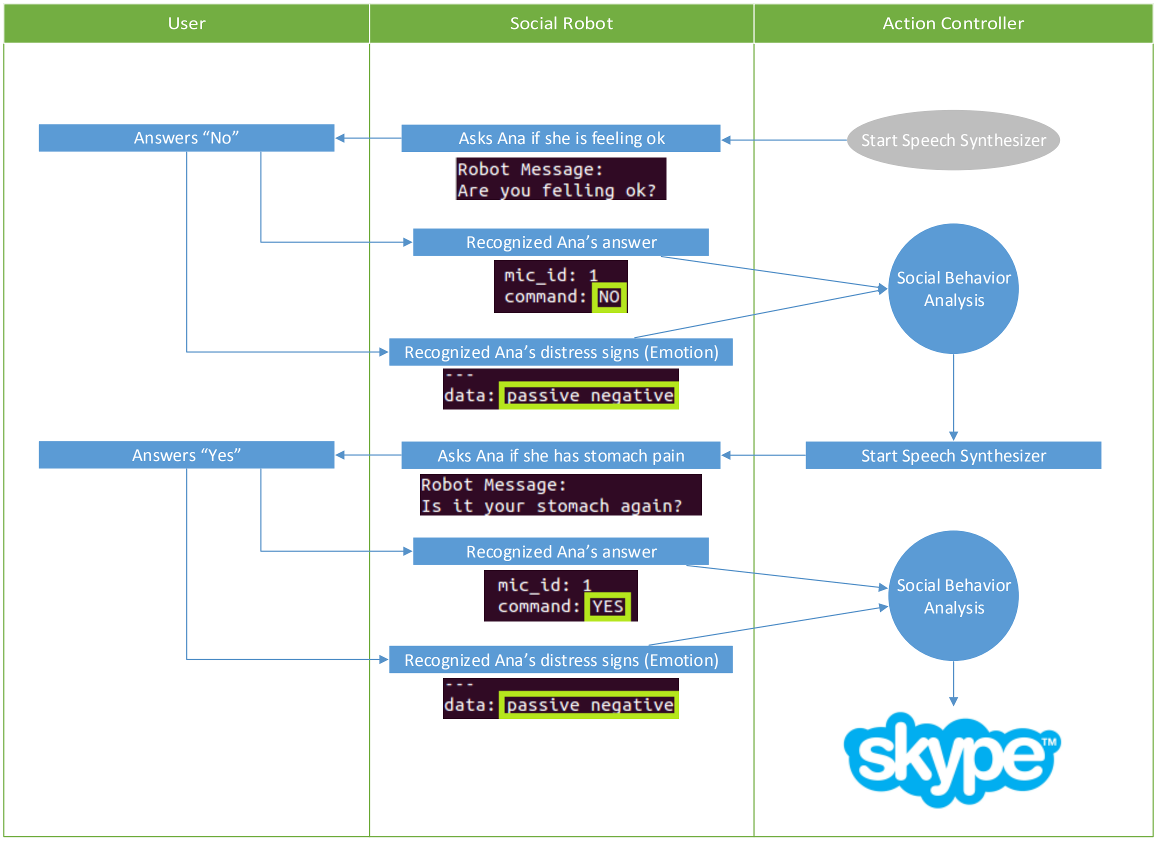
```
Robot Message:
Do you want me to request help?
```

```
mic_id: 1
command: YES
```

Figure 5.3: Simulation of the use case scenario #1.

- **Use Case Scenario 2**

This use case scenario considers the recognition of abnormal behavior. As in figure 3.4, the SocialRobot realizes that the user insists to stay in bed after her usual wake up time, starting then a conversation to retrieve information, in order to request help if necessary. As the SocialRobot is considered to be close to the user's bed and the speech recognition results are the same for each microphone, the direction of the sound source doesn't append more relevant information for the social behavior analysis, thus it was left behind for the simulation represented in figure 5.4. The user answers are also recognized with passive-negative emotional states, which combined with the speech recognition outputs lead to a video call via Skype to the care centre staff.



```
Robot Message:
Are you felling ok?
```

```
mic_id: 1
command: NO
```

```
---
data: passive negative
```

```
Robot Message:
Is it your stomach again?
```

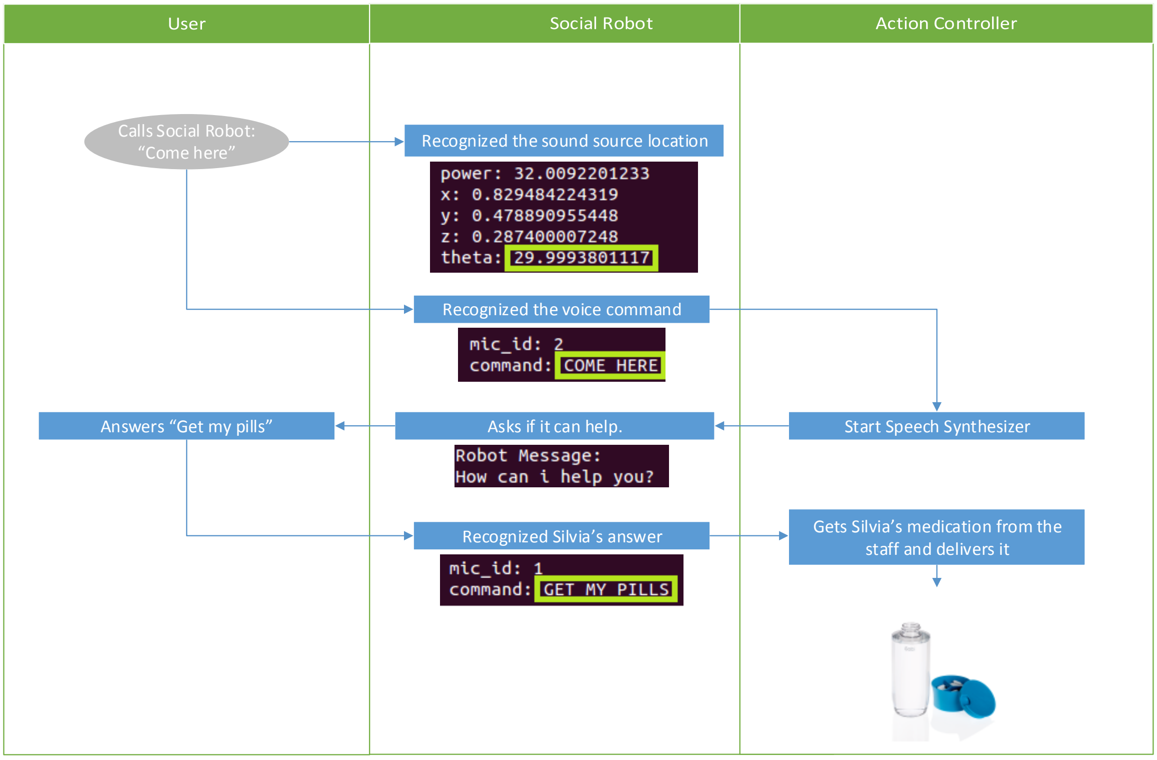
```
mic_id: 1
command: YES
```

```
---
data: passive negative
```

Figure 5.4: Simulation of the use case scenario #2.

- **Use Case Scenario 3**

The last use case scenario does not include the user's emotion for the action controller management, deciding the action by only retrieving the sound source location and speech recognition results. In this simulation, the user calls the SocialRobot, which will approach her and ask her what it could do for her, as described in figure 3.6. From figure 5.5, we can realize that the chosen microphone changes for the second command, since the Social Robot, assumes a frontal position towards the user, i.e. $\approx 0^{\text{o}}$. Since we don't consider the navigation task of the Social Robot, this simulation ends at the recognition of the second voice command.


User
Social Robot
Action Controller
Calls Social Robot: "Come here"
Recognized the sound source location
Recognized the voice command
Answers "Get my pills"
Asks if it can help.
Start Speech Synthesizer
Recognized Silvia's answer
Gets Silvia's medication from the staff and delivers it

```
power: 32.0092201233
x: 0.829484224319
y: 0.478890955448
z: 0.287400007248
theta: 29.9993801117
```

```
mic_id: 2
command: COME HERE
```

```
Robot Message:
How can i help you?
```

```
mic_id: 1
command: GET MY PILLS
```

Figure 5.5: Simulation of the use case scenario #3.

Chapter 6

Discussion and Future Work

This chapter includes a discussion of the work developed during this dissertation, in section 6.1, some ideas for future improvement of the already implemented functionalities and approaches for new capabilities, in section 6.2.

6.1 Discussion

The work conducted in this dissertation took advantage of existing toolkits for the integration of the communication system. The work mostly focus on the integration of such toolkits and the creation of ROS packages according to their features and capabilities, in order to accomplish a functional system suited to mediate interactions between the SocialRobot and its users.

Although some toolkits came as ready to use, we needed to adapt their source codes for the implementation and integration, in order to create a proper communication system. Moreover, the toolkits also cleared the way and set a reference for future research and creation of our own set of tools.

The sound capture process was realigned and modified to suit the modules bindings, accomplishing a four way classification process for the speech recognition module, only using the Pocketsphinx decoder, resulting in a final recognition result given by the best microphone according to the recognized direction (azimuth) of the sound source. Unfortunately, the emotion package couldn't exploit the same approach, due to its sound capture protocols, requiring a new device to be integrated into the communication system. Nevertheless, the emotion recognition package kept its functionality, importing the recognized emotions into the ROS environment. Moreover, the communication system can accept other emotion recognition systems, which having a more flexible audio input can restore the four way recognition process for the voice emotion package.

To realize the use case scenarios, we labeled the emotions as positive-active, positive-passive, negative-active and negative-passive accordingly to the values of arousal and valence recognized by the openEAR toolkit. Although this approach fits the requirements of the use case scenarios, using rough classes for more specific emotion may be necessary for other applications. The system can also realize those other applications, regarding more specific emotion classes, since it can endure such models.

The achieved results comprehend several tests for each modules. For the sound source localization, the range of the recognizable angles was limited from -120º to 120º degrees, since the results outside that range suffer considerable deflections, due to the Kinect's microphone enclosure and isolation. The sound source localization package only outputs the source's direction, but 3D information should be considered for future applications.

The results for the speech recognition package go through the recognition of the voice commands considered for the use case scenarios, which reduce the length of the recognizable sentences. Extending the speech recognition purposes can be perceived as the extension of the system's dictionary and language model, however, it is common to realize a descent in the recognition rates. Thus, to perform that extension, more reinforcements have to be addressed to maintain the system's robustness.

The emotion recognition package was tested using online and offline methods. For the offline tests, we used the EMO-DB database [10] to realize the package performance and recognition rates for seven emotion classes. The online tests considered our approach for the labeling of the arousal and valence values into negative and positive clusters, based the approach from Wöllmer et al. in[46]. Both recognition rates validated the toolkit and its methods, however, for more specific applications and different emotion classes, other models should be introduced regarding also idiom, cultural and demographic considerations.

Although the current integration fulfills the requirements of the use case scenarios, more complex approaches can be taken into consideration, in order to overcome some limitations and achieve more realistic scenarios. Nevertheless, this work presents a framework not only suited for elderly care services, but also for any robotic platform with human-robot-interaction prospects.

6.2 Future Work

During this master dissertation, many advances were made towards a reliable communication system, also capable of analyzing the social behavior during a conversation. However, all works regarding human-robot-interaction can be seen as a continuous research, implementation and improvement.

The future work can be divided in two parts, regarding the improvement of the already implemented capabilities and the addiction of new ones, through other approaches.

6.2.1 Improvements

- **Sound Source Localization**

The exact location of a person can be useful to acquire important clues for human-robot-interaction, such as the user's posture. Thus when someone falls and shouts for help, a 3D sound source localization can be perceived as a step forward into fall detection systems. Furthermore, the improvement of the current sound source localization will address 3D localization and distance estimation, which will enrich its role on the social behavior analysis and lead to new capabilities for the system.

- **Speech Recognition**

The speech recognition system achieved good results in the tests conducted in this dissertation, but there are still challenges ahead. For the improvement of the speech recognition capabilities of the system, two main issues should be addressed. The first one is the reliability for real-time recognition purposes, and the second should focus on the extension of the system's dictionary and language model. The voice detection algorithm could be upgraded to realize the existence of speech through probabilistic approaches, which along with noise reduction methods, would improve the system's robustness and speech recognition rates. The recognition of longer sentences should also be considered, introducing a larger dictionary and language model. This improvement will extend the purpose of the speech recognition system, from communication through voice commands into a more realistic content conversation.

- **Voice Emotion Recognition**

Emotion recognition is a complex problem with many variables to be considered. Thus, to improve the emotion recognition system and aim to more complex applications, we should consider the training of more deterministic emotion models accordingly to the environment and context presented in the scenarios of the SocialRobot project. Furthermore, the creation of our own emotion recognition system can also be considered, introducing learning capabilities through probabilistic approaches.

Considering the extension of the system's dictionary, introducing the analysis of longer speech sentences, the emotion recognition rates are also expected to improve, since the analysis of longer emotional utterances will converge into better classification outputs.

- **Action Controller**

The action controller should be considered as manager of human-robot-interactions with extensive knowledge about the users preferences and social behavior. In order to acquire such knowledge, the solution should consist in the creation and development of databases, containing relevant information about the user's social behavior and preferences, to support, improve and extend the responses of the SocialRobot. Those databases will be the action controller main asset to realize the best performable actions, according to the informations provided by the modules responsible for the user analysis. Since the action controller manages permutable tasks between the user and the system, ROS actionlib [8] functionalities can also be included to perform a better management of such tasks.

6.2.2 Beyond Auditory Analysis

Although, usually, people mostly communicate through speech, the visual cues should not be left aside. In figure 6.1, it is represented a possible visual system framework.

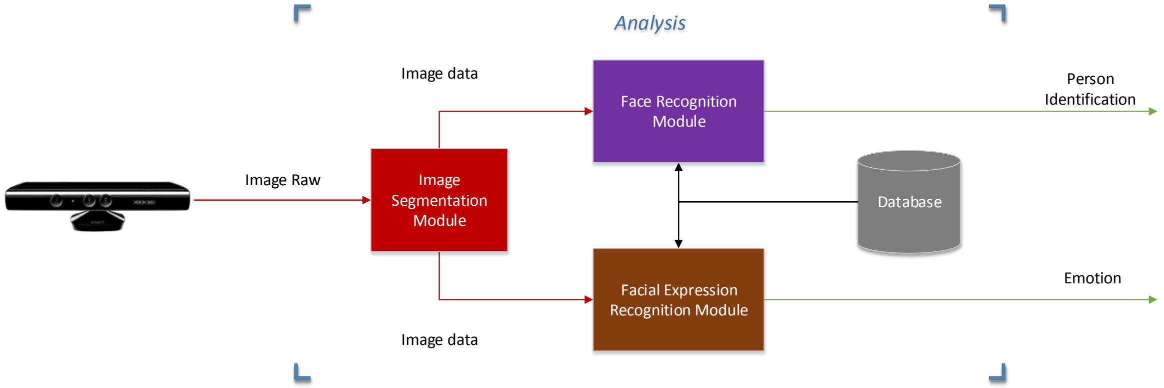


Figure 6.1: Visual system framework

The visual system consists in a preprocessing phase of image raw data capture and segmentation, followed by processing modules for face recognition and facial expression recognition purposes. Many other modules and techniques can be added to this framework, such as gesture and object recognition modules, that would complement and bring new capabilities to the communication system. Assembling audio and visual packages will give a complement of information leading to a more accurate social behavior analysis.

Chapter 7

Conclusions

The work and approaches conducted in this dissertation lead to the integration of a communication system for social behavior analysis in the SocialRobot project. Many problems were addressed during the development of this thesis, such as sound source localization, speech recognition and emotion recognition. The research of those problems and their solutions resulted in the implementation and integration of three major modules, combining all their outputs into the action controller module, which analyses the user's social behavior and manages the SocialRobot's according actions.

For the modules implementation, we used the HARK toolkit for the sound source localization solution, the Pocketsphinx decoder from the CMU-Sphinx toolkit, for the speech recognition system and the openEAR toolkit to recognize the user's emotional state. The toolkits methods were used with novel approaches, such as the four way live speech recognition process with our own voice detection algorithm and the emotion labeling for the arousal and valence values imported from the openEAR toolkit.

The integration was achieved on the Robotic Operating System (ROS), according to our novel approach for the communication system. To take advantage of ROS functionalities, we transformed the implemented modules into ROS packages with online communication capabilities. The sound source localization information provided is combined with the speech recognition results, in order to choose the best microphone and the associated recognition hypothesis for the voice commands, according to the user's location. Moreover, the best speech recognition result is also combined with the recognized emotional state of the user, in order to perform the social behavior analysis and realize the Social Robot's corresponding actions through the action controller package.

Since the Social Robot also needed to be able to communicate with the users, some complementary packages were created, introducing a speech synthesizer, an application launcher and an avatar to realize a more realistic interaction.

The assemble of all the packages conclude a communication system, capable for elderly care services, assistance and companionship, that can be integrated in any robotic platform. A social robot with communication capabilities can also be perceived as an asset for the elderly care centers, which can provide some extra help to the center's staff and improve the center's assistance and care services.

The approach presented in this work can be improved in many ways, as described in the future work addressed in the previous chapter, improving the performance of the communication system and introducing new interesting capabilities through computer vision methods and algorithms. Furthermore, the creation of our own set of tools with probabilistic approaches, introducing learning capabilities, presents interesting prospects for a future Ph.D. research.

Bibliography

[1] Cmu sphinx - speech recognition toolkit. In *http://cmusphinx.sourceforge.net*, 2013.

[2] Htk toolkit. In *http://htk.eng.cam.ac.uk*, 2013.

[3] The neuroscience of happiness. In *http://white.stanford.edu/teach/index.php/The Neuroscience of Happiness*, 2013.

[4] Pocketsphinx package. In *http://wiki.ros.org/pocketsphinx*, 2013.

[5] Robust group tutorial. In *http://www.speech.cs.cmu.edu/sphinx/tutorial.html*, 2013.

[6] Ros wiki - hark. In *http://wiki.ros.org/hark*, 2013.

[7] Social robot project webpage. In *http://mrl.isr.uc.pt/projects/socialrobot/*, 2013.

[8] ROS Wiki actionlib. In *http://wiki.ros.org/actionlib*, 2013.

[9] Paul Boersma and David Weenink. Praat: doing phonetics by computer. In *http://www.fon.hum.uva.nl/praat/*, 2013.

[10] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Interspeech*, pages 1517–1520, 2005.

[11] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[12] Ze-Jing Chuang and Chung-Hsien Wu. Emotion recognition using acoustic features and textual content. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 1, pages 53–56. IEEE, 2004.

[13] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amier, and DKJ Heylen. The sensitive artificial listner: an induction technique for generating emotionally coloured conversation. 2008.

[14] Florian Eyben, Martin Wöllmer, and Björn Schuller. Openear - introducing the munich open-source emotion and affect recognition toolkit. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.

[15] M. J F Gales and S.J. Young. Robust continuous speech recognition using parallel model combination. *Speech and Audio Processing, IEEE Transactions on*, 4(5):352–359, 1996.

[16] Michael A Goodrich and Alan C Schultz. Human-robot interaction: a survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275, 2007.

[17] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. Support vector regression for automatic recognition of spontaneous emotions in speech. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1085. IEEE, 2007.

[18] François Grondin, Dominic Létourneau, François Ferland, Vincent Rousseau, and François Michaud. The manyears open framework. *Autonomous Robots*, pages 1–16, 2013.

[19] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[20] Sue Harding, Jon Barker, and Guy J Brown. Mask estimation for missing data speech recognition based on statistics of binaural interaction. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):58–67, 2006.

[21] Toru Takahashi Ryu Takeda Keisuke Nakamura Takeshi Mizumoto Takami Yoshida Angelica Lim Takuma Otsuka Kohei Nagira Hiroshi G. Okuno, Kazuhiro Nakadai and Tatsuhiko Itohara. Hark version 1.2.0 document.

[22] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alex I Rudnicky. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[23] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech using source, system, and prosodic features. *International Journal of Speech Technology*, 15(2):265–289, 2012.

[24] Ian Lane, Vinay Prasad, Gaurav Sinha, Arlette Umuhoza, Shangyu Luo, Akshay Chandrashekaran, and Antoine Raux. Hritk: the human-robot interaction toolkit rapid development of speech-centric interactive systems in ros. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 41–44. Association for Computational Linguistics, 2012.

[25] Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. Julius—an open source real-time large vocabulary recognition engine. 2001.

[26] Chul Min Lee, Shrikanth S Narayanan, and Roberto Pieraccini. Classifying emotions in human-machine spoken dialogs. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 737–740. IEEE, 2002.

[27] Bob Mungamuru and Parham Aarabi. Enhanced sound localization. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(3):1526–1540, 2004.

[28] Kazuhiro Nakadai, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. An open source software system for robot audition hark and its evaluation. In *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, pages 561–566. IEEE, 2008.

[29] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Design and implementation of robot audition system'hark'Ñopen source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[30] T Nishiura and S Nakamura. Talker localization based on the combination of doa estimation and statistical sound source identification with microphone array. In *Statistical Signal Processing, 2003 IEEE Workshop on*, pages 597–600. IEEE, 2003.

[31] Cátia Pinho, João Filipe Ferreira, Pierre Bessière, Jorge Dias, et al. A bayesian binaural system for 3d sound-source localisation. In *International Conference on Cognitive Systems (CogSys 2008)*, 2008.

[32] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[33] Social Robot Project. Specification of user needs and requirements. In *Deliverable D1.1*, 2012.

[34] Social Robot Project. Use case scenarios development and ethical and privacy considerations. In *Deliverable D1.2*, 2012.

[35] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[36] Ben Rudzyn, Waleed Kadous, and Claude Sammut. Real time robot audition system incorporating both 3d sound source localisation and voice characterisation. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 4733–4738. IEEE, 2007.

[37] Alberto Sanchis, Alfons Juan, and Enrique Vidal. A word-based naive bayes classifier for confidence estimation in speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):565–574, 2012.

[38] Björn Schuller, Ronald Müller, Florian Eyben, Jürgen Gast, Benedikt Hörnler, Martin Wöllmer, Gerhard Rigoll, Anja Höthker, and Hitoshi Konosu. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760–1774, 2009.

[39] Björn Schuller, Zixing Zhang, Felix Weninger, and Felix Burkhardt. Synthesized speech for model training in cross-corpus recognition of human emotion. *International Journal of Speech Technology*, 15(3):313–323, 2012.

[40] Liang-Che Sun and Lin shan Lee. Modulation spectrum equalization for improved robust speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):828–843, 2012.

[41] Yuki Tamai, Satoshi Kagami, Yutaka Amemiya, Yoko Sasaki, Hiroshi Mizoguchi, and Tachio Takano. Circular microphone array for robot's audition. In *Sensors, 2004. Proceedings of IEEE*, pages 565–570. IEEE, 2004.

[42] Hesham Tolba and Douglas O'Shaughnessy. Speech recognition by intelligent machines. *IEEE Canadian Review*, 38:20–23, 2001.

[43] Keith Vertanen. Baseline wsj acoustic models for htk and sphinx: Training recipes and recognition experiments. *Cavendish Laboratory, University of Cambridge*, 2006.

[44] N. Virag. Single channel speech enhancement based on masking properties of the human auditory system. *Speech and Audio Processing, IEEE Transactions on*, 7(2):126–137, 1999.

[45] Volker Willert, Julian Eggert, Jürgen Adamy, Raphael Stahl, and E Korner. A probabilistic model for binaural sound localization. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(5):982–994, 2006.

[46] Martin Wöllmer, Florian Eyben, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. Data-driven clustering in emotional space for affect recognition using discriminatively trained lstm networks. In *INTERSPEECH*, pages 1595–1598, 2009.

[47] Phillip C Woodland, Julian J Odell, Valtcho Valtchev, and Steve J Young. Large vocabulary continuous speech recognition using htk. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 2, pages II–125. IEEE, 1994.

[48] Takeshi Yamada, Satoshi Nakamura, and Kiyohiro Shikano. Robust speech recognition with speaker localization by a microphone array. In *Spoken Language, 1996.*

ICSLP 96. Proceedings., Fourth International Conference on, volume 3, pages 1317–1320. IEEE, 1996.

[49] Cha Zhang, Dinei Florêncio, Demba E Ba, and Zhengyou Zhang. Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. *Multimedia, IEEE Transactions on*, 10(3):538–548, 2008.

[50] Zhengyou Zhang. Microsoft kinect sensor and its effect. *Multimedia, IEEE*, 19(2):4–10, 2012.