Intelligent Speech Interaction with Mechatronic System using HMM

Apostolos Tsagaris\(^1\), Konstantinos-Hercules Kokkinidis\(^2\) and Athanasios Manitsaris\(^3\)

\(^1,2,3\)University of Macedonia, Department of Applied Informatics, MTCG Lab, Thessaloniki, Greece, e-mail: ‘tsagaris@uom.edu.gr, "kkokk@uom.gr, ‘manits@uom.edu.gr

Abstract— This paper proposes an interaction with mechatronic system based on voice commands. The goal of this research is to develop a human-machine interface that could be able to control a mechatronic system by using voice commands. With the use of low features extracted from the voice commands, the system will be trained to recognize specific commands. After the training, the commands are recognized and transferred to the controlled system. Intelligent techniques are used with a combination of Hidden Markov Models and Dynamic Time Wrapping for the training and the recognition. The LEGO MINDSTORMS NXT robotics platform controlled by custom software developed under max/msp program is the evaluation tool of the system. The system is tested with different set of commands and the experimental results show that the interaction is efficient with a number of 92\% of accuracy on voice commands.

Index Term— Human machine Interaction, Voice control, Intelligent control, mechatronic systems

I. INTRODUCTION

Nowadays, the evolution of technology leads to an increasing development of new applications of intelligent control of machines and robots. It is necessary to find natural ways of interaction between human and machines, to be easier to communicate ordinary people without special knowledge or robots. In this direction, many scientific fields are collaborating such as artificial intelligence, human-machine interaction, machine vision, biology and psychology.

The research areas that deal mainly with communication of human with machine use knowledge afforded by the sectors dealing with humans and how they act and behave and try to discover new methods and new technologies for human machine interaction. So the development of technology and science has given the opportunity to develop systems and tools for process automation and effective communication of people with computers and general machinery. Some of those are using speech. System such as call centers, customer services, information services, alternative safeguards, smart homes and more. The main concern of the speech system should be friendly to the user, because of direct communication with humans. The speech technology is in the field of science that deals with the communication between man and machine (HCI) through speech. Also interested in the harmonization of speech with other forms of communication such as face recognition, emotions, gestures and more. The speech is the most important way of communication between humans, characterized by specific features that cannot be easily captured in writing and give another dimension to the transmitted information. A lot of specific requirements can be identified on voice recognition technology for controlling a robotic or mechatronic system. These are noise robustness, high accuracy, low hardware requirements, or the capability to run on specific hardware and software configurations. In the most cases the researcher are using specific words to control the system.

In this paper, it is presented a methodology for the real-time voice command control of mechatronic system. There are three main parts. The speech recognition and feature extraction, the machine learning technique based on Hidden Markov Models (HMM) and Dynamic Time Wrapping (DTW) and the control of mechatronic – robotic system. The paper consists of six sections. The first section is an introduction in Speech Recognition and Features, Hidden Markov Models (HMM) and Dynamic Time Wrapping (DTW) and in Mechatronics systems. In the second section is described the state of the art and in the third the methodology overview. The forth section consists of the system architecture and the fifth describe the experimental results. Finally in the conclusions and the future work are presented in section six.

A. Speech recognition and features

Speech recognition is the technology that captures words spoken by a human with a help of microphone and recognized them by speech recognizer. The process of speech recognition consists of different steps. An ideal situation in the process of speech recognition is that a speech recognition system recognizes all words performed by a human. This is practically difficult because the performance of a speech recognition system depends on a large number of factors. Features, users, noise, e.t.c are some of the factors. This research deals with the issue of feature selection and extraction.

The features that are used in this research are selected based on the literature review and are Spectral Centroid, Spectral Flux, Spectral Rolloff, Short Time Energy and Zero Crossing Rate.

Spectral Centroid

The spectral centroid is the center of gravity of the magnitude spectrum of the short time Fourier transforms (STFT)

$$C_t = \frac{\sum_{i=1}^{N} M_f[i] \cdot n}{\sum_{i=1}^{N} M_f[i]}$$  \hspace{1cm} (1)

Where \( M_f[i] \) is the magnitude of the Fourier transform at frame \( t \) and frequency bin \( n \) [1]

Spectral Flux

The spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions
\[ F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \]  
\[ \sum_{n=1}^{R_{2}} M_{t}[n] = 0.85 \times \sum_{n=1}^{N} M_{t}[n] \]

**Spectral Rolloff**

The spectral rolloff is defined as the frequency \( R_x \) below which 85% of the magnitude distribution is concentrated [1]

**Short Time Energy (STE)**

The short-time energy of an audio signal is defined as

\[ E_t = \frac{1}{2} \sum_{m} [x(m) - w(n-m)]^2 \]

where \( x(m) \) is the discrete time audio signal, \( n \) is time index of the short-time energy, and \( w(m) \) is a rectangle window of length \( N \). It provides a convenient representation of the amplitude variation over time [2].

**Zero Crossing Rate (ZCR)**

The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back [3]. The short-time average zero-crossing rate, gives rough estimates of spectral properties of audio signals.

\[ Z_n = \frac{1}{2} \sum_{m} [s_{ gn}[x(m)] - s_{ gn}[x(m-1)]] \times w(n-m) \]

where

\[ s_{ gn}[x(nm)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \]

and \( w(m) \) is a rectangle window [2].

**B. Hidden Markov Models (HMM) and Dynamic Time Wrapping (DTW)**

HMMs are chosen to classify the voice commands, and their parameters are learned from the training data. With the help of the most likely performance criterion, the voice commands can be recognized by evaluating the trained HMMs.

Because the HMM is more feasible than the Markov model, we adopt the former to learn and recognize the continuous hand gestures to direct robots.

HMM classifier: The last phase is the classification of command. The probability that each HMM has produced the input vector constitutes a recognition criterion. Each HMM models a movement. Thus, the HMM which has the highest probability of having created the input sequence corresponds to the most probable voice command represented in this input vector.

An HMM can be determined by:

- \( \{S\} \) – Set of states, included initial state SI and final state SF,
- \( T \) – Array of transition probabilities, \( T = \{t_{ij}\} \), \( t_{ij} \) is the transition probability from the state I to state j,
- \( O \) – Array of output probabilities, \( O = \{o_j(x)\} \) for continuous HMM.

Forward HMM is using to model the dynamic features of each voice command. A model is created for each command and the parameters of models were appreciated by the sequences of training that were available for each command, using the algorithm Baum-Welch. During the decoding phase, each HMM is able to produce the sequence of observation. This probability constitutes a recognition criterion. The sequence is seen as belonging to the gesture of which HMM gives the highest probability. If a sequence gives too low probabilities (smaller than a threshold) in all models, then this sequence is not considered as belonging in any command which our system recognizes.

Thus it allow us, with the proper training of the HMM model to predict the command during the speaking. It is not necessary to complete the command. With the help of hidden Markov Model the command will be predicted before the end of the speaking. Finally, the predicted command will be transferred to the mechatronic system to control it automatically. Because of the different duration of each execution of the appropriate command at the same time the Dynamic time Wrapping algorithm it is used to synchronize the two time series (training and recognition) (Figure 1).
This method calculates the similarity between two sequences which can differ in time or in speed [4].

C. Mechatronic systems

Mechatronics is the integration of electro – mechanical and computer technologies into complex products (Figure 2). It is a combination of precision mechanical engineering, electronic, control engineering and computers for the intelligent control of machines.

A mechatronic system is a combination of actuators, controllers and sensors. Typical mechatronic applications are transportation systems, power generation systems, CNC machines, etc. The most typical example of mechatronic systems is the robotics. Robotics is the art of design and construction of reprogrammable elements - devices flexible and able to perform different functions. The level of automation is much more flexible and shows future trends in mechatronics.

A good example of mechatronic-robotic system is a LEGO system. Solutions such as LEGO are often used, in which we can quite easily have a simulation of the system. The LEGO MINDSTORMS NXT robotics system consists of servo motors and a wide variety of digital and analog sensors. Designed for easy assembly, it is nonetheless powerful enough to create sophisticated robots and teach advanced concepts like closed-loop control systems and embedded systems. It is a platform for the development of practical works with mechatronic systems. It has several advantages. The first one is the flexibility that this system has, since due to its sensors and actuators, and with the help of a programmable control unit, it is possible to develop a great variety of projects and activities. Another advantage is the price: with the construction pieces and the electronic devices this platform has a very low cost, which is much more economical than the price of other robotic platforms. The system allows the communications by USB and Bluetooth. In this way, the limitations of the infrareds are eliminated since much more longer distances are allowed and they don’t need a correct orientation between the sender and the receiver and moreover, the communications are not lost if there is an obstacle between them.

II. STATE OF THE ART

There have been several studies on the interaction of human and robots. There are many applications in which the user interacts with the system in various ways. Examples are control applications with gestures, nods, finger movements, etc. [5], [6]. Loper et al. [7] present a robot that follows people indoors and outdoors using a depth camera for recognizing static gestures. The case of Alex Couture-Beil [8] gives an example of control of multi-robotic systems through gestures. A user by combining facial and gesture recognition controls more than one robotic system simultaneously. Looking to a side of the system, the system is activated and controlled by the gesture [8]. D.Riek et al. [9] and Iba et al. [10] propose a direct communication through collaborative gestures using humanoid robot and found satisfactory results in recognizing gestures. Wai Shan Ng; Sharlin, E. [11] use gestures to control flying robotic systems. The gestures associated with the use of a racket for better motion detection. Another field of application is the field of medicine with the various management information systems in the health sector. For example the use of system interaction with gestures to navigate user data in CT and MRI scanners [12], [13], [14].

The field of interaction with voice commands is also a great research. St. Ondas et. al. [15] proposes an automatic speech recognition system, which enables to control functions of the system using a set of voice commands and is based on Hidden Markov model. The system was designed to give information in both direction. Human to robot and Robot to human. According to the authors the presented modules were organized in the multimodal interface for controlling functions of the robotic system which were working well in laboratory conditions. However, in order to evaluate the usability of the interface and to find possible problems, it was necessary to make tests in real conditions. In Poncela’s and Gallardo-Estrella’s work [16] it is presented a Human Robot Interface to teleoperate a robotic platform via the user’s voice. With the help of a speech recognition a user-dependent acoustic model has been developed to teleoperate a robot with a set of commands. The experimental results have been successful, both in terms of a high recognition rate and the navigation of the robot under the control of the user’s voice. Barkana et. al. [17] proposes a speech recognition technique to include patient’s feedback into a robot-assisted habilitation system, with an accuracy more than 90%.

In the field of machine learning technique for voice event classification there are several approaches. Usage of Support Vector Machines (SVMs) [18], GMM binary trees [19] and Hidden Markov models (HMMs) [20] are the most popular. Berger et al present a comparison between HMM and Dynamic Bayesian Networks (DBN) and have shown some advantages of the DBN framework (higher flexibility, intuitiveness in model design, etc) as well as some drawbacks (mainly, its potential complexity) [22]. Kayikci et al. [23] utilized Hidden Markov Models and a neural associative memory for learning to understand short speech commands in a three-staged recognition procedure.

III. METHODOLOGY’S OVERVIEW

In this way, the aim of the methodology is to identify effective voice commands. First are setting voice control commands and assigned to appropriate movements of the robot. Thus it is like a dictionary of movements, which directs...
the order of the robot that each movement must perform. Then as recorded by suitable recording device (microphone), voice commands are processed by the system to extract the necessary features. The features that are used in the proposed methodology are short-time energy, spectral centroid, spectral flux, spectral rolloff and zero-crossing rate. With the help of these features the system is trained with machine learning techniques to recognize each command (Figure 3). The technique which is used in both training and recognition is a combination of Hidden Markov Models (HMM) and Dynamic Time Wrapping (DTW).

![Diagram of system overview](image)

Fig. 3. Methodology’s overview

In the next step there will be recognized the voice command, which actually comprises the values of their characteristics. With the help of the extracted feature data the system is trained with HMM to recognize the commands. The recognized command is transferring to the robotic system and the robot executes the specific command.

Once the system is trained then the voice commands are performed by users and classified according to the appropriate features. Finally the predicted commands will be transferred automatically to the Lego Mindstorm mechatronic system to control it via wireless communication.

The advantages of the wireless communication, with the mechatronic system allow the continuously sending and receiving of information to execute commands in real time [21].

The evaluation of the system is performed by cross-validation Precision and Recall techniques. Cross-Validation is useful for overcoming the problem of over-fitting. Over-fitting is a term which refers to when the model requires more information than the data can provide. It is one of the most commonly used model selection criteria. Based on a data splitting, part of the data is used for fitting each competing model (or procedure) and the rest of the data is used to measure the performance of the models.

Precision-recall method is a parameterized method that is balanced between accuracy and noise. Recall is the number of correctly recognized commands by the system divided by the total number of commands (i.e. recognized and not recognized commands). Precision is the ratio of correctly recognized commands to the total number of commands recognized by the system (i.e. correctly recognized and wrongly recognized commands). Recall measures how sensitive the recognized recognition module is and precision determines the accuracy of the system’s module.

In these terms, precision is the probability that the recognition’s event is valid, and recall is the probability that the ground truth data was recognized. Equations (6) and (7) provide mathematical definitions of precision (p) and recall (r) for convenience.

**Precision** is the fraction of recognized commands that are relevant to the search.

\[
\text{Precision} = \frac{\text{Relevant \_commands} \cap \text{Recognized \_commands}}{\text{Recognized \_commands}} \tag{6}
\]

**Recall** in information retrieval is the fraction of the strokes that are relevant to the query that are successfully retrieved.

\[
\text{Recall} = \frac{\text{Relevant \_commands} \cap \text{Recognized \_commands}}{\text{Relevant \_commands}} \tag{7}
\]

IV. SYSTEM OVERVIEW

The complete system consists of three subsystems, the speech recording system, the control platform and the robot. The control platform is implemented under Max/MSP programmable language. It is a programming tool that allows the user to create programs graphically and is concentrated in multimedia development, focusing primarily in the music field. The developed software works either in real-time by recording and processing the speech commands directly from the input, or with pre-recorded files. The main patch gives to user the opportunity to record a voice command or to execute the feature extraction in real time (Figure 3).
The feature will be extracted, and the intelligent system will be trained (Figure 5) with the feature vector described in the methodology.

Once the system is trained to recognize all the pre-selected commands, the user may give a specific command and the system after the recognition sends it to the robot. The robot will execute the specific command.

V. EXPERIMENTAL RESULTS

Speech commands are recorded in mono channel as 48000 samples per second and 16 bits per sample. The total duration of each command is approximately 3.2 sec with a number of total samples close to 150,000 per command. As speech is low bandwidth, these values are quite sufficient. As the window size, different values are tested and shown that 512 is the best window size (hamming) for this project.

There are 5 fixed commands which can be used to control basic robot actions. “Go Forward” is the command which moves the robot straight forward, “Go Backward” is the command which moves the robot straight backward, “Go Right” is the command which turns the robot right, “Go Left” is the command which turns the robot left and “Stop” is the command which stops the robot.

The set of commands will be executed by the user 9 times. For the evaluation of the system it is used cross-validation (Leave one out). From the 9 sets of commands the first set is used to train the system and the other 8 sets are used for the recognition. The recognition efficiency is recorded. The second set of commands is used for training and the other 8 (1, 3-8) sets are used for the recognition process. This will continue till the end of the sets. Totally 72 executions are performed for each command. The recognition efficiency is shown in Table I.

![Fig. 4. Voice Commands evaluation Software](image)

![Fig. 5. Intelligent system for voice recognition](image)

Table I

<table>
<thead>
<tr>
<th>Recognition sets</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>Not</th>
<th>Prec.</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>59</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
<td>88.1</td>
<td>92.2</td>
</tr>
<tr>
<td>C2</td>
<td>2</td>
<td>62</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>91.2</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>3</td>
<td>1</td>
<td>61</td>
<td>2</td>
<td>1</td>
<td>89.7</td>
<td>93.8</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>63</td>
<td>5</td>
<td>94</td>
<td>92.6</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>60</td>
<td>7</td>
<td>92.3</td>
<td>89.6</td>
</tr>
</tbody>
</table>

For the first command execution the system recognize the command 1 59 times like command 1 (the correct), 4 times like command 2, 2 times like command 3, 2 times like command 4 and 5 times like nothing. The recognition for the other commands is showing in Table I. Precision and Recall methodology produced the outputs presented in Table I. Totally there are given 72 sets of voice commands. Precision is the percentage of accurate commands which is recognized according to the total of commands, while Recall is the percentage of commands which is related to the total actual commands. The total precision of the system in 360 voice commands is 91.06% and the total recall in 360 voice commands is 92.44%

VI. CONCLUSION AND FUTURE WORK

An automatic speech recognition system is proposed, which is based on a combination of Hidden Markov models and dynamic time wrapping and enables to control functions of a system using a set of voice commands. The importance of command recognition lies in building efficient human–machine interaction. The algorithm is searching for the right command and with the help of hidden Markov Model and dynamic time wrapping the command can be predicted before the end of the speech. The command will be transferred to the mechatronic system to control it automatically. The described architecture is simple and economical, features low computational cost, high reliability and high robustness to noise and working conditions. The interaction is extremely simple and natural and does not require the user to use any other additional device.

The efficiency of the system is very high with a precision rate between 88% to 92% and a recall rate between 90% to 94% for the 5 commands, and an average value from 91.06% for precision and 92.44% for recall.

Although the proposed set of commands to teleoperate the robot is fixed, one of the advantages of the system is its easy adaptation to new commands and new platforms. In future research the methodology can be used to implement a complete control system which can be combined with gestures. Furthermore, in future efforts the use of language model can also be added so as the system will be able to take orders not just according to its training but to a wider language model used in common PCs.
REFERENCES


