Anadolu Üniversitesi Bilim ve Teknoloji Dergisi A- Uygulamalı Bilimler ve Mühendislik
Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering

2017 - Volume: 18 Number: 5
Page: 897 - 907
DOI: 10.18038/aubtda.346121
Received: 27 October 2016 Revised: 10 May 2017 Accepted: 24 June 2017

REAL-TIME CONTROL OF MOBILE ROBOT USING HMM-BASED SPEECH RECOGNITION SYSTEM

Hayrettin TOYLAN [1,*], Erol TÜRKEŞ [2], Evren ÇAĞLARER [1]

[1] Department of Mechatronics Engineering, Technology Faculty, Kırklareli University, Kırklareli, Turkey
[2] Department of Mechanical Engineering, Engineering Faculty, Kırklareli University, Kırklareli, Turkey

ABSTRACT

Human-robot interaction (HRI) is a significant area of interest in robotics which has attracted a wide variety of studies in recent years. In order to provide natural human-robot interaction, robots will have to acquire the skills to detect and to integrate meaningfully information from multiple modalities. In this paper, a practical speech-controlled mobile robot car system is presented and discussed. In this study the developed Hidden Markov Model (HMM) with separate word recognition system and real-time control were obtained on a mobile robot. Mel-Frequency Cepstral Coefficients (MFCC) were applied as features for the control design of mobile robot. In the study, 270 speech commands (İLERİ=forward, GERİ=backward, DUR=stop, SAĞA=right, SOLA=left) which are collected from 54 different people were applied to a series of mathematical operations and 12 cepstral coefficients were derived. Therefore, a database was generated by 12 cepstral coefficients. Thus, HMM model was trained and tested according to database. Speech data were classified in two groups as 90% training data and 10% test data. The recognition success rate of test commands was measured 94%.

Keywords: Hidden markov model, MFCC, Speech recognition, Mobile robot, Robot

1. INTRODUCTION

Human robot interaction (HRI) is a multidisciplinary field with the contributions from human computer interaction, artificial intelligence, robotics, natural language understanding, design, and social sciences. Today, robots are often described as artificial agents with detectable capacity by researchers. The use of robots is most widespread for serial and automatic production in factories. Today, however, the use of robots is becoming increasingly widespread in all types of workplaces and in critical workspaces. This is further accelerated by the increasing use of robotics and robot learning. Thanks to the development of technology and science, humans and robots can interact to make them useful in more useful and critical tasks. Robot movements are also determined by robot detection in human robot communication. This is known as sensory motor coordination [1, 2]. Until today, many researchers have searched for social relationships between humans and robots. For example, Kismet was developed for studying early caregiver-infant interaction [3]. Also, a robot that stands in a line [4] and a robot that talks with multiple people [5] have been developed. Furthermore, various communicative behaviors using a robot's body have been discovered, such as a joint-attention mechanism [6]. Also, one of the most important issues of HRI studied by researchers is speech recognition. Speech recognition is a technology where the system understands the words (not its meaning) given through speech. Speech synthesis is the key technologies for HRI systems, and the interactive robot is one of the most typical and important applications to be realized in HRI systems. Communication between man and machine should be basically written or verbal. That is, a robot should be able to receive written or verbal commands. There must be a natural interaction between human and robot. It also provides the ability to transfer information while receiving direct feedback during the interaction. The humankind communicates with robots using different dialects and vocabulary. Therefore, robots should know the language of the person

*Corresponding Author: hayrettintoylan@klu.edu.tr

who is in communication and learn new words. Even the robots should know the dialect of the person who is in contact. In other words, robots should have the ability to speech recognition. Speech recognition is the process of automatically recognizing spoken words of a person based on information in the voice signal during speech. Recognition technique makes it possible to the speaker's voice to be used in verifying their identity and control access to services. The most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC [7, 9]. Today, automatic speech recognition systems are considered as potential computation models of these cognitive skills. Despite the huge advances of automatic speech recognition in the last decade, all conventional automatic speech recognition systems still perform substantially worse than humans [10, 11]. Today, the availability of large speech databases has increased. Thus, by applying statistical learning algorithms, it is possible to construct speech synthesis systems called database or corpus based approach. These systems can be trained automatically. In addition, these systems not only produce natural and high-quality synthetic speech, but can also produce audio features of the original speaker. In order to create such a system, hidden Markov Models (HMMs) have increased in popularity. HMMs are successfully applied for modeling the sequence of speech spectra in speech recognition systems. The performance of HMM based speech recognition systems has been improved by techniques that use the flexibility of HMMs. These techniques are context dependent modeling, dynamic property parameters and mixture of Gaussian densities, binding mechanism, speaker and environment adaptation techniques. Many speech synthesis systems can synthesize high quality speech. However, speech synthesis systems still cannot synthesize speech, including voice features, such as speech styles, emotions. In speech synthesis systems based on the selection and combination of acoustic units, a large amount of speech data is required to obtain various audio features. However, collecting these speech data is rather difficult and time consuming. An HMM based speech synthesis system is proposed to generate speech synthesis systems that can produce various voice features [12, 13].

In this paper, a practical speech controlled mobile robot car system is proposed. The proposed speech controlled mobile robot car simultaneously processes speech, integrates perceptual models for robot audition, thus it moves in the direction of the speech command. Speech recognition process of mobile robot car was created in real time with the help of the Matlab program without using the toolkit. In addition to this, instead of a word based speech recognition, a phoneme based speech recognition system is used. The authors have introduced a Speech Controlled Robot (SCR) based on Mel frequency cepstral coefficient (MFCC) and Hidden Markov model (HMM). The system works in several systematic steps; Design of mobile robot car, Signal pre-processing, Feature extraction, Speech Recognition. When the operations specified in this work are done in sequence, Human–robot interaction will be provided via speech commands.

2. MATERIALS AND METHODS

2.1. Design of Mobile Robot Car

Speech Controlled Robot (SCR) is a mobile robot whose motions can be directed by the user by giving speech command. Therefore, pertinent electric motors take action and SCR moves according to speech command. Mechanic drive system of the SCR is formed available two gear-boxes with two wheels. To move of gear-boxes used in two DC motors as showed in Figure 1.

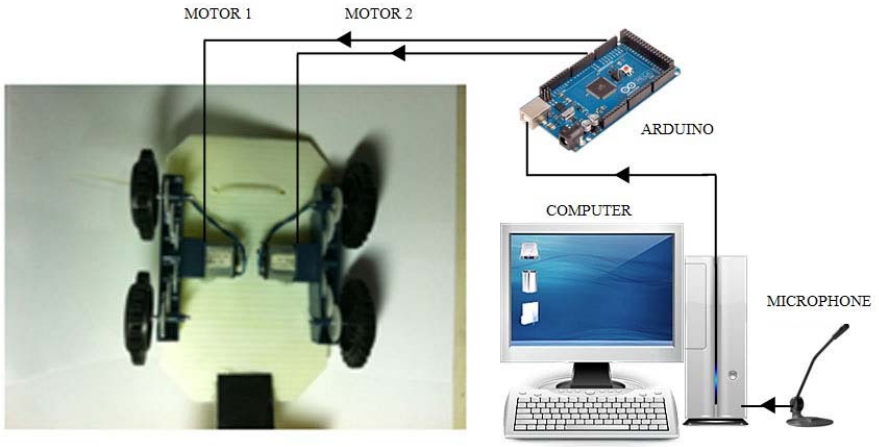

Figure 1. Mechanic drive system of the SCR

The software of the speech command is designed by using Matlab. Giving speech commands to PC microphone are processed by Matlab. According to the results of speech command algorithm, electrical signals are transmitted to DC motors via Arduino Mega 2560 used as DAQ card. Motion forms of DC motors, according to the defined commands, are given by Table 1.

Table 1 State of the DC Motors

Turkish commands in this study	**DC Motor 1**	**DC Motor 2**
İLERİ	Forward	Forward
GERİ	Back	Back
DUR	Stop	Stop
SOLA	Stop	Forward
SAĞA	Forward	Stop

Design and software of SCR are created to perform 90 degree rotation. In the rotation direction, one of the motors will be stopped during rotation, other motor will be moved in determined working time. Thus, SCR turns to intended direction.

2.2. Signal pre-Processing

Speech command recognition process starts with the determination of command limits and features which are the best way to represent command progress. In this stage, the speech command signals are prepared for smooth playback, followed by high-pass filtration and segmentation. During signal pre-processing, the following steps were included [14]:

Filtering: To exceed a certain energy threshold of the signal from the user indicating the start of the audio signal. Drops below a certain energy threshold indicate that the audio signal has been terminated. Speech command signals were filtered using a high-pass filter to remove unwanted low-frequency components. To prevent unwanted noise from the outside passing audio data is made available to the segmentation.

Segmentation: Segmentation operations are applied to determine the commands in the area where speech data recorded after the filtering process. Islets occurs in the speech data after the filtering process. The largest of these islands constitutes our speech command signal. Other islands are deleted.

2.3. Feature Extraction

The feature extraction is the calculation of a set of feature vectors that provides a compact representation of the speech signal. There are a few feature extraction methods to use in speech recognition. Today, features that are commonly used as MFCC (Mel Frequency Cepstral Coefficients) which represents the human voice in a better way [15]. In this study, The Mel Frequency Cepstral Coefficients (MFCC) algorithm was used. This algorithm is presented in Figure 2 block diagram.

Figure 2. Procedure of MFCC

Speech signal is divided into frames, each having the same period. These frames intersect in certain areas. Then all of the frames are passed from windowing algorithms called as Hamming Windowing. Thus, the calculation of coefficients has characteristics divided into frames and windowing signals facilitate and thus, a continuous signal is obtained. In this study, the audio signals were separated into frames of 25 ms in length, overlapping 15 ms. And each frame 25 ms length was windowed by a Hamming Windowing (Figure 3). Sampling rate is 44100 Hz.

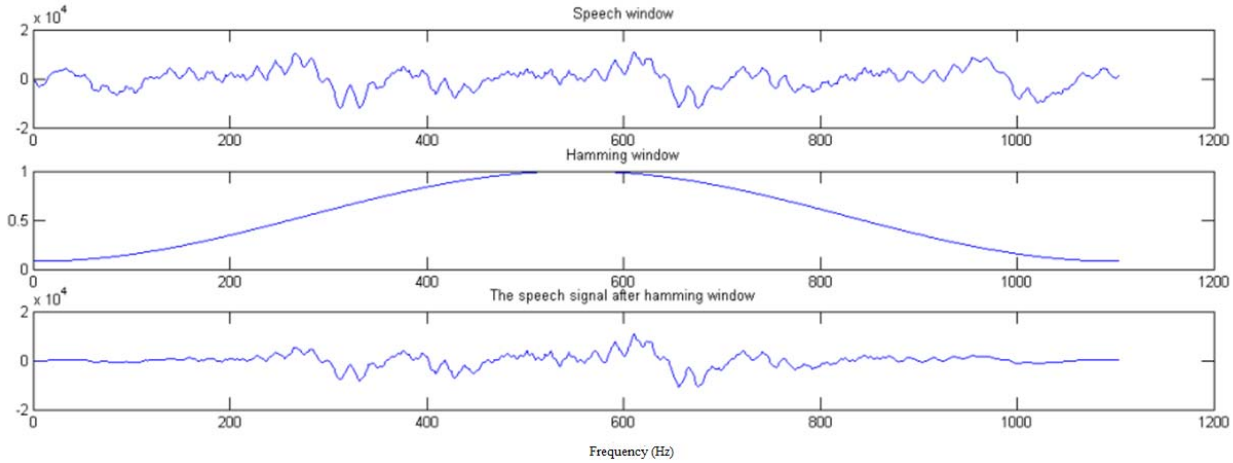

Figure 3. (a) Speech frame **(b)** 25 ms Hamming Window **(c)** Marked status of Speech signal

Mathematical expression of Hamming Windowing is given as [16];

$$w(n) = 0.54 - 0.46\cos\left[2\pi n/(N-1)\right] \qquad N-1 \geq n \geq 0 \tag{1}$$

In the next step, amplitude spectrum of windowing signal is obtained by Fast Fourier Transform (FFT). A Fast Fourier Transform (FFT) is applied to convert the speech segment of N samples from the time domain to the frequency domain. Mathematical expression for N pieces data is given as [16];

$$X(k) = \sum_{J=1}^{N} x(j) w_N^{(j-1)(k-1)} \tag{2}$$

where

$$w_N = e^{(-2\pi i)/N} \tag{3}$$

Mel scale is linearly up to 1 kHz, after 1 kHz it is expressed logarithmically on a scale at varying intervals. Given f (Hz) frequency in the following equation should be used to express the frequency scale [14];

$$f_{mel} = 2595 \log_{10}(1 + f_{linear} / 700) \tag{4}$$

Sign calculated amplitude spectrum is passed from the mel scale filter bank in a next step. The Mel scale filter bank consists of linear filters up to 1 kHz, in case higher than 1 kHz frequencies, it consists of logarithmically placed triangle filters. In Figure 4 is shown mel scale filter bank. The signal is passed through a mel scale filter set located at a frequency range of 0-22050 Hz.

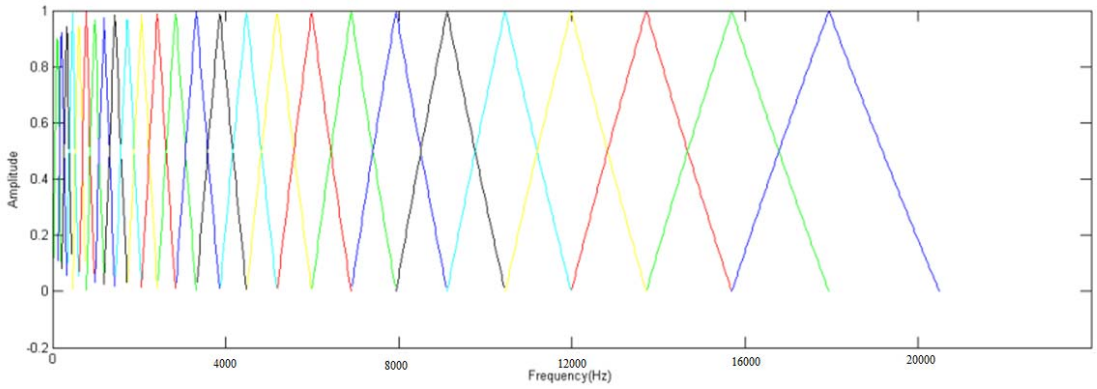

Figure 4. Mel-scale filter bank

After the logarithm of the amplitude spectrum was obtained, passing back to the time setting with discrete cosine transform coefficients MFCC 12 was obtained. Thus, a filter much more similar to human ear is obtained. The number of features have been increased to 36 with the first and second derivatives of the 12 features.

2.4. Speech Recognition

5 types of speech commands collected from 54 different people was determined as "forward, reverse, stop, turn right and turn left". These various types of speech commands were used for HMM (Hidden Markov model) modeling. HMM is one of the successful techniques for acoustic modeling due to its analytical ability in the speech phenomenon and its accuracy in practical speech recognition systems. Five models of speech command were trained individually by HMM and unknown input speech command signals were classified automatically by the trained models. HMM model parameters are estimated in the training phase by maximum likelihood based using training data sets [17, 18].

HMM models used in the application are necessary to solve three basic problems. These problems; what is the observation likelihood, to estimate most likely hidden case sequence and how to re-adjust the model parameters.

In this study, sequence likelihood of HMM is observed with forward algorithm. In an HMM model given $\lambda = (\pi, A, B)$ parameters and for observation series $O= O_1, O_2, O_3,.. O_T$ of this model $P(O \mid \lambda)$ the probability of the observation sequence forward algorithm $\alpha_t(i)$,

$$\alpha_t(i)=P(O_1O_2.....O_t,q_t=S_i|\lambda) \tag{5}$$

According to the λ model, $\alpha_t(i)$ that forward variable is probability of the $O_1,O_2..O_T$ partial observation sequence at S_i situation in t time. $\alpha_t(i)$ variable is solved by a inductive method as given below [19, 20].

i. Starting-Giving the first value

$$\begin{aligned} \alpha_1(i) &= P(O_1, q_1 = S_i \mid \lambda) = P(O_1 \mid q_1 = S_i, \lambda)P(q_1 = S_i, \lambda) \\ \alpha_1(i) &= \pi_i b_i(O_1), \qquad t=1,\ 1 \le i \le N \end{aligned} \tag{6}$$

ii. Iteration

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad 1 \le t \le T-1, 1 \le j \le N \tag{7}$$

iii. Termination

$$P(O_1 \mid \lambda) = \sum_{i=1}^{N} P(O, q_t = S_i \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{8}$$

This step gives as the sum of the latter feed forward variable of $P(O \mid \lambda)$ account required

$$\alpha_t(i) = P(O_1O_2.....O_t, q_t = S_i \mid \lambda) \tag{9}$$

The next step for O_1, O_2, O_3, O_T.. observation sequence and λ model, is to select the case sequence $Q = q_1\ q_2\ q_3$ which explains these observations in the most appropriate form. A formal method depends on the dynamic programming technique of Viterbi Algorithm used to find the case sequence. Finally, we show how to change the model parameters $\lambda = (\pi, A, B)$ to maximize $P(O \mid \lambda)$ probability. The Baum-Welch algorithm is used in predicting the parameter $\lambda = (\pi, A, B)$ to maximize the probability of $P(O \mid \lambda)$ locally.

3. RESULTS

Firstly, 5 different commands were obtained from 54 different people to use in mobile robot application. Each sample phrase is taken from one person, a total of 270 speech data were collected (Table 2).

Table 2. Some of the system parameters

Parameter	Value
Sampling rate	44100Hz,16 bits
Database	Isolated 5 Turkish words
Total Data	270
Speakers	54
Window type	Hamming

Records taken for the training and testing were pretreated. High-pass filter is applied to the obtained speech data to prevent the unwanted low-frequency signs. The data over a certain threshold are extracted. After applying a high pass filter to audio files, the findings are shown in red (Figure 5). The filtering process is completed by adding the opposite sign of the obtained data as a result of the high pass filtering process.

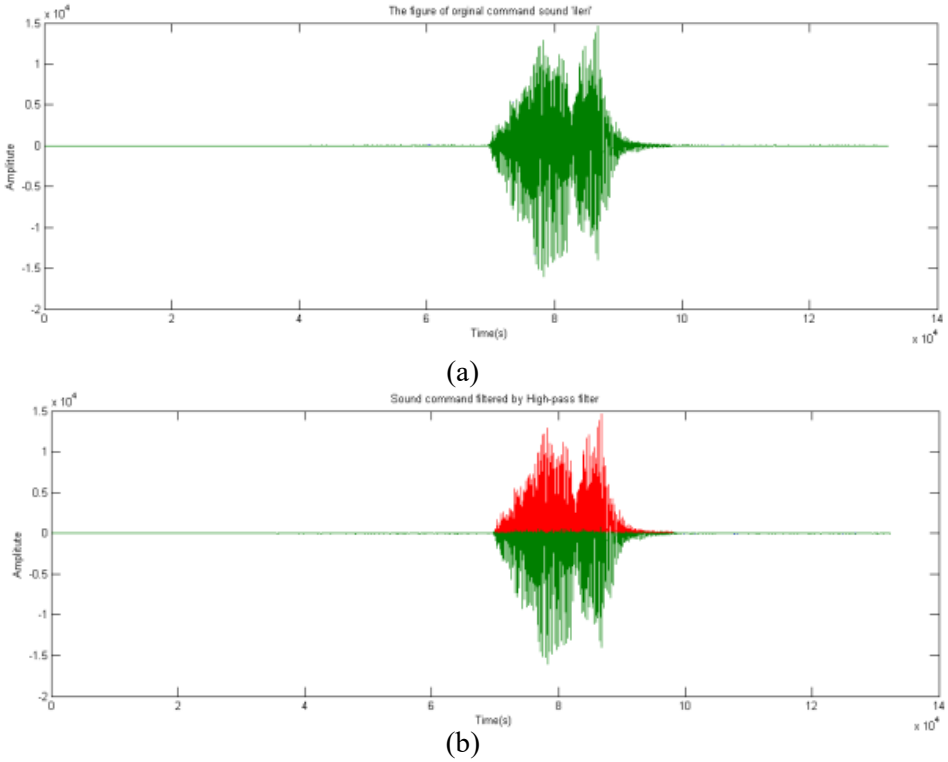

Figure 5. (a) Original speech command **(b)** The speech command prepared high-pass filter

After the filtering process; the segmentation of the speech data is performed. The segmentation of the filtered speech data may cause multiple islands to occur. The largest island is defined as the desired speech command. The obtained result of filtering and segmentation processes speech command is shown in Figure 6.

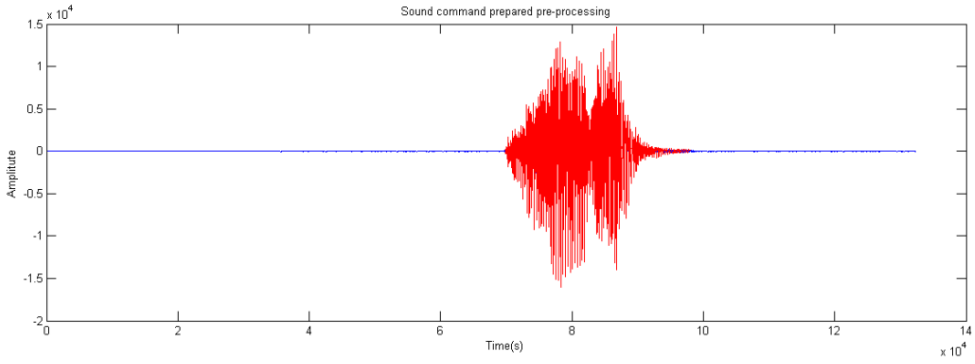

Figure 6. The speech command after prepared pre-processing

After determining the boundaries of speech commands, voice features are obtained and stored in the training database. In this study, the 36 features are used for determining the characteristics of speech data. MFCC which based on FFT (Fast Fourier Transform) is a numerical analysis method used to mimic the perception of the human ear. The MFCC is much less affected from the changes and structure of the sound waves according to the other methods. The algorithm uses 12 features of the MFC confident. The response to a speech command is the 12 different MFC that is shown in Figure 7. Also, 1st and 2 nd derivatives of MFC confidents are used as a feature and the number of features is determined as 36 in this study. Confidents stored in the database is obtained and stored to be used later in the testing phase.

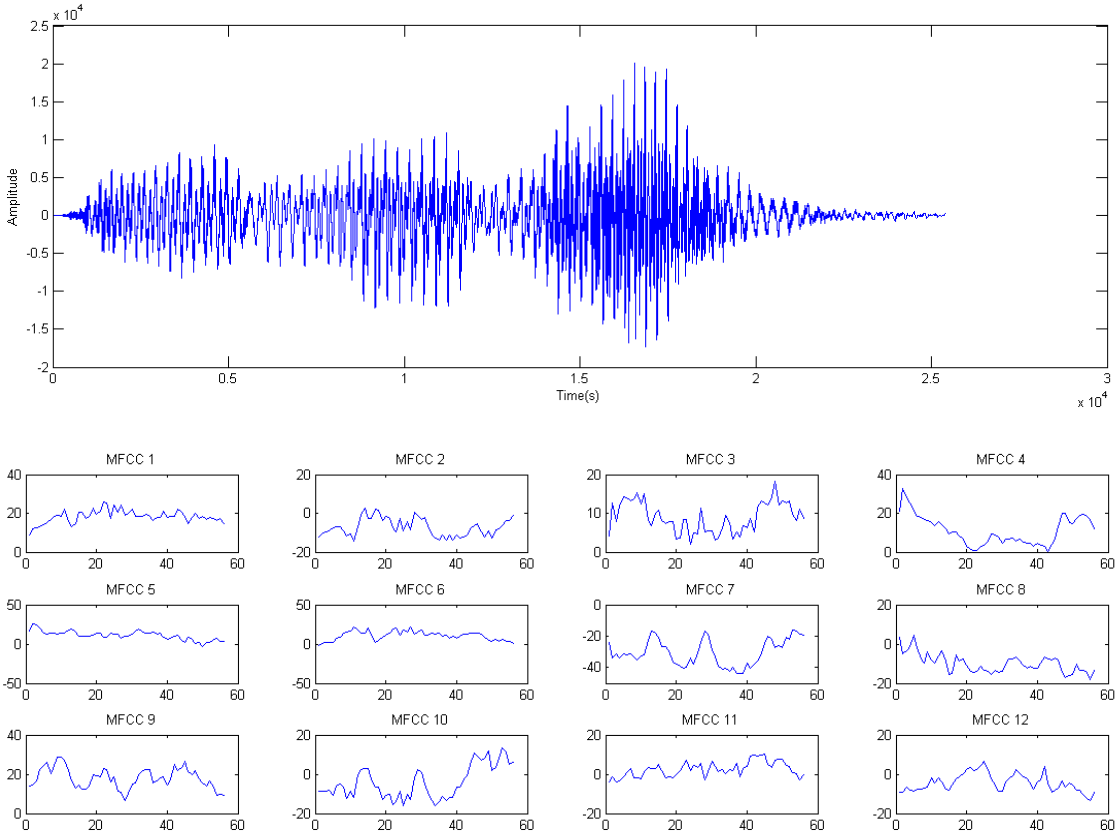

Figure 7. Speech command before MFCC and (b) MFCC coefficients

Classification with high accuracy of speech commands requires an efficient feature extraction and classification. The command recognition system consists of two schemes. The first phase of the model is the training phase, the second is the testing phase. HMM model parameters $\lambda = (\pi, A, B)$ are calculated for each speaker in the training phase of the training data. The model parameters are obtained by calculating the Baum-Welch (Forward-Backward) algorithm. The number of cases which used in the training phase of the system is chosen as 15. These values have been selected on the basis of experience. In the testing phase, the test data was applied to the input of the command recognition system and Viterbi algorithm is used to define the most likely sequence of cases.

This makes the model maximum likelihood and determines which of the test data is on command in the database. Speech data were classified in two groups as 90% training data and 10% test data. The data used for training phase was not used in the test phase. Recognition success rate, obtained by test data was found as 94%. In Table 3 is shown used parameters and test results.

Table 3. Speech recognition test results

Parameter	Value
MFCC	12
Models Training & Recognition	HMM
Training data	243
Test data	27
Success rate	%94

4. DISCUSSION AND CONCLUSIONS

In this study is carried out application of human robot interaction with the help of real-time speech recognition algorithm. When mobile robot system is run, the speech command automatically is analyzed in every three seconds in order to control the mobile robot. When speech-controlled mobile robot car system (SCR) is run, the speech command automatically is analyzed in every three seconds in order to control the SCR. The system asks you to say a command by directing users and the movement of the SCR is provided by obtained a new data from the command recognition. In this study, the speech command recognition algorithm based on HMM and MFCC has been developed successfully in order to select the right speech command of the mobile robot. In this paper, three basic problems of HMM algorithm are solved to be used real-world application. For solution of the first problem, HMM model parameters $\lambda = (\pi, A, B)$ are calculated using Baum-Welch (Forward-Backward) algorithm. For solution of the second problem, Viterbi algorithm is used to define the most likely sequence of hidden cases and the lastly maximum likelihood model is used to re-adjust the model parameters.

A phoneme based speech recognition system is used instead of a word based speech recognition. Phoneme numbers and similar phonemes used in word affect speech recognition success rate. There are 29 phonemes in Turkish alphabet but the amount of sound is much more than due to different accent. This is an important condition to collect the data. The speech recognition success rate of the selected words was found as 94%. It was observed that the recognition success of the words was affected by environmental noises. In order to increase the performance of the speech recognition success rate, it is possible to perform operations such as noise reduction by using different software before processing speech data to the recognition algorithm.

Mobility of the improved SCR has 5 different positions. Because of the system is trainable, angular rotations can be achieved by increasing the number of instructions and the different functions can be added to the system. For future studies it may be suggested that to practice isolated speech or continuous speech with the length of medium word or larger word. And also it can be experienced for Turkish language.

REFERENCES

[1] Pfeifer R, Bongard J. How the body shapes the way we think: A new view of intelligence. London, England: MIT Press, 2006.

[2] Schillaci G, Verena VH. Prerequisites for Intuitive Interaction - on the example of Humanoid Motor Babbling. HRI 2011 Workshop on Expectations in intuitive human-robot interaction; 6 March 2011; Lausanne, Switzerland. pp. 23-27.

[3] Breazeal C, Scassellati B. A context-dependent attention system for a social robot. IJCAI '99 Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence; 31 July- 6 August 1999; Stockholm, Sweden: IEEE. pp. 1146-1151.

[4] Nakauchi Y, Simmons R. A Social Robot that Stands in Line, Auton. Robot 2002; 12: 313-324.

[5] Nakadai K, Hidai K, Mizoguchi H, Okuno HG, Kitano H. Real-Time Auditory and Visual Multiple-Object Tracking for Robots. International Joint Conference on Artificial Intelligence; 4-10 August 2001; Washington, USA: IEEE. pp. 1425-1432.

[6] Scasscllati B. Investigating Models of Social Development Using a Humanoid Robot, Bio robotics. Virgina, ABD: MIT Press, 2000.

[7] Ahmed Q. Al-thahab, Control of Mobile Robot using Speech Recognition, Journal of Babylon University, Pure and Applied Sciences, No. (3), Volume 19 : 2011.

[8] Chauhan G, Chaudhari P, Robotic Control using Speech Recognition and Android, International Journal of Engineering Research and General Science, 2015, Volume 3, Issue 1, pp. 1210-1216.

[9] Razuri JG, Sundgren D, Rahmani R. Speech Emotion Recognition in Emotional Feedback for Human-Robot Interaction, International Journal of Advanced Research in Artificial Intelligence, 2015, Vol. 4, No.2, pp. 20-27.

[10] Mubin O, Bartneck C, Feijs L, Huysduynen HH, Hu J, Muelver J. Improving Speech Recognition with the Robot Interaction Language, Disruptıve Science and Technology, 2012, Volume 1, Number 2, pp. 79-88.

[11] Clemente IA, Heckmann M, Wrede B. Incremental word learning: Efficient HMM initialization and large margin discriminative adaptation. Speech Commun 2012; 54: 1029–1048.

[12] Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T. Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis. IEEE International Conference on Acoustics, Speech and Signal Processing; 5-9 June 2000; Istanbul, Turkey: IEEE. pp. 1315-1318.

[13] Tokuda K, Zen H, Black AW. An HMM-based speech synthesis system applied to English. Proceedings of 2002 IEEE Workshop on; 11-13 Sept. 2002; Santa Monica, USA:IEEE. pp. 227-230.

[14] Chauhan S, Wang P, Lima CS, Anantharaman V. A computer-aided MFCC-based HMM system for automatic auscultation. Comput Bio. Med 2008; 38: 221 – 233.

[15] Martinez J, Perez H, Escamilla E, Suzuki MM. Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) Techniques.22nd International Conference on Electronics Communications and Computing; 12 February 2012; Cholula, Mexico: pp.248 - 251.

[16] Pramanik A, Raha R. Automatic Speech Recognition using correlation analysis, Information and Communication Technologies (WICT); 30 October -2 November 2012; Trivandrum, India: IEEE. pp. 670-674.

[17] Najkar N, Razzazi F, Sameti H. A novel approach to HMM-based speech recognition systems using particle swarm optimization. Math. Comput. Model 2010; 52: 1910–1920.

[18] Rabiner LR. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE1989; 77 (2): 257–286.

[19] Lawrence R, Juang BH. Fundamentals of Speech Recognition. New Jersey, USA: Prentice Hall International Inc., 1993.

[20] Can T, Öz E. Saklı Markov modelleri kullanılarak Türkiye'de dolar kurundaki değişimin tahmin edilmesi. Istanbul U J Sch Bus Admin 2009; 38: 1-23.