An Approach For: Improving Voice Command processor Based On Better Features and Classifiers Selection

Mohamed FEZARI¹, Ali Al-Dhioud²

1. Laboratory of Automatic and Signals, Annaba, BP. 12, Annaba, 23000, ALGERIA
2. Faculty of IT Al-Zaytoonah University, Amman, Jordan
Mohamed.fezari@uwe.ac.uk ,aldahoud@alzaytoonah.edu.j

Abstract: Details of designing a voice command system based on combining features and pipelining classifiers is presented in this work. Based on research and experimental results, more features will increase the rate of recognition in ASR (automatic Speech recognition). Thus combining classical components used in automatic speech recognition system such as Crossing Zero, energy, MFCC with wavelet transform (to extract meaningful formants parameters) followed by a pipelining ordered classifiers GMM and HMM has contributed in reducing the error rate considerably. To implement the approach on a real-time application, a Personal Computer interface was designed to control the movements of a four degree of freedom robot arm by transmitting the orders via radio frequency circuits. The voice command system for the robot arm is designed and tests showed an Improvement by combining techniques.

Keywords: Speech recognition, wavelet transform, GMM/HMM, pipelining and hybrid techniques, Human-machine Communication, wireless command.

1. Introduction

Robot arms, or manipulators, comprise a 2 billion dollar industry. Bolted at its shoulder to a specific position in the assembly line, the robot arm can move with great speed and accuracy to perform repetitive tasks such as spot welding and painting. In the electronics industry, manipulators place surface-mounted components with superhuman precision, making the portable telephone and laptop computer possible. Teleoperated Robot arm are widely used in space robotics, bomb disposal, and remotely operated vehicles. [1] and [2].

Yet, for all of their successes, these commercial robots suffer from a fundamental disadvantage: lack of human voice control. Human-robot voice interface is an interesting application for elderly people and impaired users. A fixed manipulator has a limited range of commands provided by a manipulator. Mainly using a joystick, keyboard or a mouse.

This paper proposes a new voice command system (VCS) for teleoperating a robot arm. Based on the recognition of isolated words, using a set of traditional pattern recognition approaches and a discrimination approach based pipelining classical methods [2][5] and [7] in order to increase the rate of recognition. The increase in complexity as compared to the use of only traditional approach is negligible, but the system achieves considerable improvement in the matching phase, thus facilitating the final decision and reducing the number of errors in decision taken by the VCS.

It is only recently that the fields of robot arm control and AGV (Automatic Guided Vehicle) navigation have started to import some of the existing techniques developed in AI for dealing with non-stationary information.

Hybrid method is a simple, robust technique developed to allow the grouping of some basic techniques advantages. It therefore increases the rate of recognition. The selected features are: Zero Crossing and Extremes (CZEXM), continuous Wavelet transform CWT specially Gabor wavelets, Energy Segments (ES), and cepstral coefficients (MFCC: Mel Frequency Cepstral coefficients).[6]. The application uses a set of commands in Arabic words in order to control the directions of four DOF robot arm. It has to be implemented on programmable integrated circuits [8] and has to be robust to noise.

As application, a voice command for a four DOF robot arm is chosen. The robot arm is “TERGANE - TR45”. There have been many research projects dealing with robot control, among these projects, there are some projects that build intelligent systems [10-12]. Voice command system needs the recognition of isolated words from a limited vocabulary used in guiding a Robot Arm Control system (RACS)[14]. The paper is presented as follow: in section 2, we present the selected feature computation, in section 3 the speech recognition agent is detailed based on GMM and HMM, the hardware part is presented in section 4, finally in section 5 and 6 we present the results of tests simulation and a conclusion.

2. Gabor Wavelets and Speech Analysis

In recent years, the wavelet transform has been successfully applied to image and signal processing. Unlike the Fourier transform, the wavelet transform can detect the characteristics of local signals.
Therefore, the wavelet transform is particularly helpful for the analysis and processing of speech signal where local formants information is important.

Gabor wavelets, which are similar to the Mexican hat wavelets provide a more powerful tool for analyzing speech and music. We shall first go over their definition, and then illustrate their use with some examples. A Gabor wavelet, with width parameter $w$ and frequency parameter $v$, is the following analyzing wavelet:

$$\Psi(x) = \omega^{-1/2} e^{-\pi (x/w)^2 + i^2 m^2 x/w}$$

This wavelet is complex valued. Its real part $R(x)$ and imaginary part $I(x)$ are.

$$\Psi_R(x) = w^{-1/2} e^{-\pi (x/w)^2} \cos(2\pi v x / w)$$

$$\Psi_I(x) = w^{-1/2} e^{-\pi (x/w)^2} \sin(2\pi v x / w)$$

The width parameter $w$ plays the same role as for the Mexican hat wavelet; it controls the width of the region over which most of the energy of ($x$) is concentrated. The value $v/w$ is called the base frequency for a Gabor CWT.

One advantage that Gabor wavelets have when analyzing sound signals is that they contain factors of cosines and sines, as shown in (2.2) and (2.3). These cosine and sine factors allow the Gabor wavelets to create easily interpretable scalograms of those signals which are combinations of cosines and sines—the most common instances of such signals are recorded music and speech. More details on this is in [6-8], but first we need to say a little more about the CWT defined by a Gabor analyzing wavelet.

Because a Gabor wavelet is complex valued, it produces a complex-valued CWT. For many signals, it is often sufficient to just examine the magnitudes of the Gabor CWT values and take them as features. CWT were used to select some principles features within the uttered signal that can not be extracted by classical techniques these components will indeed increase the true positive counter however they will add on calculation period.

![Figure 1](image.png)

Figure 1. a-Mexican hat CWT (scalogram) of a test signal with two main frequencies, b- Magnitudes of Gabor scalogram of test signal.

3. Hybride Techniques

Regarding the different ASR paradigms proposed during these years, Hidden Markov Models of Gaussian mixtures (HMM/GMM) [19] is doubtless the most widely accepted approach.

Alternatively, Artificial Neural Networks (ANN) based framework has also been proposed [2], but despite their high discrimination ability in short-time classification tasks, they have proved inefficient when dealing with long-term speech segments. With the scope of solving the problem of long time modeling of the ANN framework, one of the most successful alternatives to HMM/GMM was later proposed, commonly known as hybrid ANN/HMM or connectionist paradigm [20]. In general, hybrid architectures seek to integrate ANN ability for estimation of Bayesian posterior probabilities into a classical HMM structure that permit modeling long-term speech evolution.

3.1 GMM technique

The GMM can be viewed as a hybrid between parametric and non-parametric density models. Like a parametric model, it has structure and parameters that control the behavior of density in known ways. Like non-parametric model it has many degrees of freedom to allow arbitrary density modeling. The GMM density is defined as weighted sum of Gaussian densities:

$$P_{G,M}(x) = \sum W_m g(x, m, C_m)$$

the total number of Gaussian components. $W_m$ are the component probabilities ($\_w = 1$), also called weights.

We consider K-dimensional densities, so the argument is a vector $x = (x_1, \ldots, x_K)^T$. The component probability density function (pdf), $g(x, m, C_m)$, is a K-dimensional Gaussian probability density function (pdf) given in Equation 7 as follows:

$$g(x, m, C_m) = e^{-1/2(x - \mu_m)^T}$$

Where $\mu_m$ is the mean vector, and $C_m$ is the covariance matrix.

Organizing the data for input to the GMM is important since the components of GMM play a vital role in making the word models. For this purpose, we use K-means clustering technique to break the data into 256 cluster centroids. These centroids are then grouped into sets of 32 and then passed into each component of GMM. As a result, we obtain a set of 8 components of GMM. Once the component inputs are decided, the GMM modeling can be implemented more details are in [21].
3.2 HMM basics

Over the past years, Hidden Markov Models have been widely applied in several models like pattern [6], or speech recognition [6][9]. To use a HMM, we need a training phase and a test phase. For the training stage, we usually work with the Baum-Welch algorithm to estimate the parameters (π, A, B) for the HMM [7, 9]. This method is based on the maximum likelihood criterion. To compute the most probable state sequence, the Viterbi algorithm is the most suitable.

A HMM model is basically a stochastic finite state automaton, which generates an observation string, that is, the sequence of observation vectors, \( O = O_1, O_2, ..., O_T \). Thus, a HMM model consists of a number of N states \( S = \{ S_i \} \) and of the observation string produced as a result of emitting a vector \( O_t \) for each successive transition from one state \( S_i \) to a state \( S_j \). \( O_t \) is d dimension and in the discrete case takes its values in a library of M symbols.

The state transition probability distribution between state \( S_i \) to \( S_j \) is \( A = \{ a_{ij} \} \), and the observation probability distribution of emitting any vector \( O_t \) at state \( S_j \) is given by \( B = \{ b_{j}(O_t) \} \). The probability of distribution initial state is \( \pi = \{ \pi_i \} \).

\[
a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \quad (3.1)
\]

\[
B = \{ b_{j}(O_t) \}. \quad (3.2)
\]

\[
\pi, = P(q_0 = S_i) \quad (3.3)
\]

Given an observation O and a HMM model \( \lambda = (A, B, \pi) \), the probability of the observed sequence by the forward-backward procedure \( P(O|\lambda) \) can be computed [10]. Consequently, the forward variable is defined as the probability of the partial observation sequence \( O_1O_2,...,O_t \) (until time t) and the state S at time t, with the model \( \lambda \) as \( \alpha(i) \), and the backward variable is defined as the probability of the partial observation sequence from t+1 to the end, given state S at time t and the model \( \lambda \) as \( \beta(i) \). The probability of the observation sequence is computed as follow:

\[
P(O|\lambda) = \sum_{i=1}^{N} \alpha_i(i)^* \beta_i(i) = \sum_{i=1}^{N} \alpha_i(i) \quad (3.4)
\]

and the probability of being in state I at time t, given the observation sequence \( O \) and the model \( \lambda \) is computed as follow:

\[
\pi_t = P(q_0 = S_t) \quad (3.5)
\]

4. Proposed technique: Pipelining Recognition System Models

The main elements are shown in the block diagram of Figure 1. The pre-processing block is used to adapt the characteristics of the input signal to the recognition system. It is essentially a set of filters, whose task is to enhance the characteristics of the speech signal and minimize the effects of the background noise produced by the external conditions and the motor.

![Figure 2.b: BLOCK DIAGRAM OF THE PROPOSED RECOGNITION SYSTEM MODELS](image)

The SL implemented is based on analysis of crossing zero points and energy of the signal, the linear prediction mean square error computation helps in limiting the beginning and the end of a word; this makes it computationally quite simple.

The parameter extraction block analyses the signal, extracting a set of parameters with which to perform the recognition process. First, the signal is analysed as a block, the signal is analysed over 20ms frames, at 256 samples per frame. Five types of parameters are extracted: Normalized Extremes Rate with Normalized Zero Crossing Rate (CZEM), Gabor wavelets coefficients, Energy Segments (ES) and MFCC’s [4]. These parameters were chosen for computational simplicity reasons (CZEM, ES), robustness to background noise (12 Cepstral parameters) and robustness to speaker rhythm variation (DTWE) [7]. Gabor wavelets for extracting useful parameters which are formants.

The reference pattern block is created during the training phase of the application, where the user is asked to enter 50 times each command word. For each word and based on the ten repetitions, ten vectors of parameters are extracted from each segment and stored. The matching block, by application of GMM classifier then HMM models in order, it compares the reference patterns and those extracted from the input signal in both methods. The matching decision integrate: a hybrid recognition block based on five methods, and a weighting vector. The value of this vector is then taken from the results of these methods. Example: if GMM and then HMM recognise the first word in table 1 then decision block will recognise the first word as uttered.

Tests were made using each method GMM and HMM separately. From the results obtained, a weighting vector is extracted based on the rate of recognition for each method. Figure 1 shows the elements making up the main blocks for the hybrid recognition system (HRS).

A voice command system and an interface to the robot arm are implemented. The voice commands consists of Arabic words used in mobile robot control
5. Experiments on the System

The corpus is based on 50 students from the 1st and 2nd year Master Degree of our department, they have already studied a course on robotics and have a lecture on Human-Machine Communication. 40 students were used for training phases and 10 students were used for tests.

The developed system has been tested within the laboratory of L.A.S.A. The tests were done only on the five command words. Three different conditions were tested:

1. The rate of recognition using the classical methods with classical parameters used in ASR (ES, MFCC) and GMM as classifier.
2. The rate of recognition using the classical methods with classical parameters used in ASR (ES, MFCC) and HMM as classifier.
3. The rate of recognition with enhancement of parameters used in ASR (ES, MFCC, CWT) and GMM as classifier.
4. The rate of recognition with enhancement of parameters used in ASR (ES, MFCC, CWT) and HMM as classifier.
5. The rate of recognition with enhancement of parameters used in ASR (ES, MFCC, CWT) and combining GMM and HMM as classifier.
6. The effect of computation time due to combination of different techniques.

For the four first tests, each command word is uttered 50 times. The recognition rate for each word is presented in Table 1.

Fig.4.4 presents the improvement using combined features and classifiers. Fig. 4.b, represents the computation time for each combination (hybridizing classifiers), it is clear that the computation time is long by using more data parameters to process or more classifiers in pipeline to select.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Words</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Fourth</th>
<th>Fifth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ammame</td>
<td>60</td>
<td>70</td>
<td>72</td>
<td>75</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>Wara</td>
<td>65</td>
<td>71</td>
<td>74</td>
<td>82</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>Yamine</td>
<td>70</td>
<td>75</td>
<td>75</td>
<td>76</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Yassar</td>
<td>68</td>
<td>72</td>
<td>76</td>
<td>76</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Kif</td>
<td>75</td>
<td>78</td>
<td>80</td>
<td>84</td>
<td>92</td>
<td></td>
</tr>
</tbody>
</table>

6. Conclusion And Future Work

Evaluations show the interest of teleoperating arm manipulator for users' even handicapped people. Since the designed robot consists of a microcontroller and other low-cost components namely RF transmitters, the hardware design can easily be carried out. The results of the tests shows that a better recognition rate can be achieved inside the laboratory and especially if the phonemes of the selected word for voice command are quite different. However, a good position of the microphone and additional filtering may enhance the recognition rate. Several interesting applications of the proposed system different from previous ones are possible as mentioned in section 5.

Beside the designed application, a hybrid approach to the implementation of an isolated word recognition agent HSR was used. This approach can be implemented easily within a DSP or a CMOS DSP microcontroller or even FPGA modules.

The use of hybrid technique based on classical recognition methods makes it easier to separate the class represented by the various words, thus simplifying the task of the final decision block. Tests carried out have shown an improvement in performance, in terms of misclassification of the words pronounced by the user. Segmentation of the word in three principal frames for the Zero Crossing and Extremes method gives better results in recognition rate. Another line of research will be in the future relates to natural language recognition based on semantic and large data base.

References


