Voice Control for a Gripper Using Mel-Frequency Cepstral Coefficients and Gaussian Mixture Models

Velasco-Hernandez, Gustavo [1]. Díaz-Toro, Andrés Alejandro [2]

Perception and Intelligent Systems Research Group
School of Electric and Electronics Engineering
Universidad del Valle

Abstract

This work presents an implementation of a speaker-dependent speech recognition system used to control a gripper. The application was made using MATLAB and the gripper was assembled using the Lego Mindstorm NXT robotic kit. Four commands are implemented for controlling the gripper: Open, close, rotate left and rotate right. The development was divided into two stages. In training stage, we use Mel Frequency Cepstral Coefficients (MFCCs) and Gaussian Mixture Models (GMMs) to generate a representation of each defined command. Then, in testing stage, those models are used to identify the speakers utterance and send the command to the actuator. Finally, we present test results that show a performance of 95.09% for our system, and then we compare it with similar works.

Methodology

The methodology developed along this work is based on two main features. The first one is that the system is focused on detecting isolated words; this means that the speaker will say words separated by silence spaces. The second feature is that the system detects words only for the speaker for whom the models were computed. The development of this system was divided into training stage (figure 1) and development stage (figure 2). Training stage is intended to build acoustic model based on the utterances of speakers. Testing stage capture speech signal and use previously created models to validate it and execute commands.

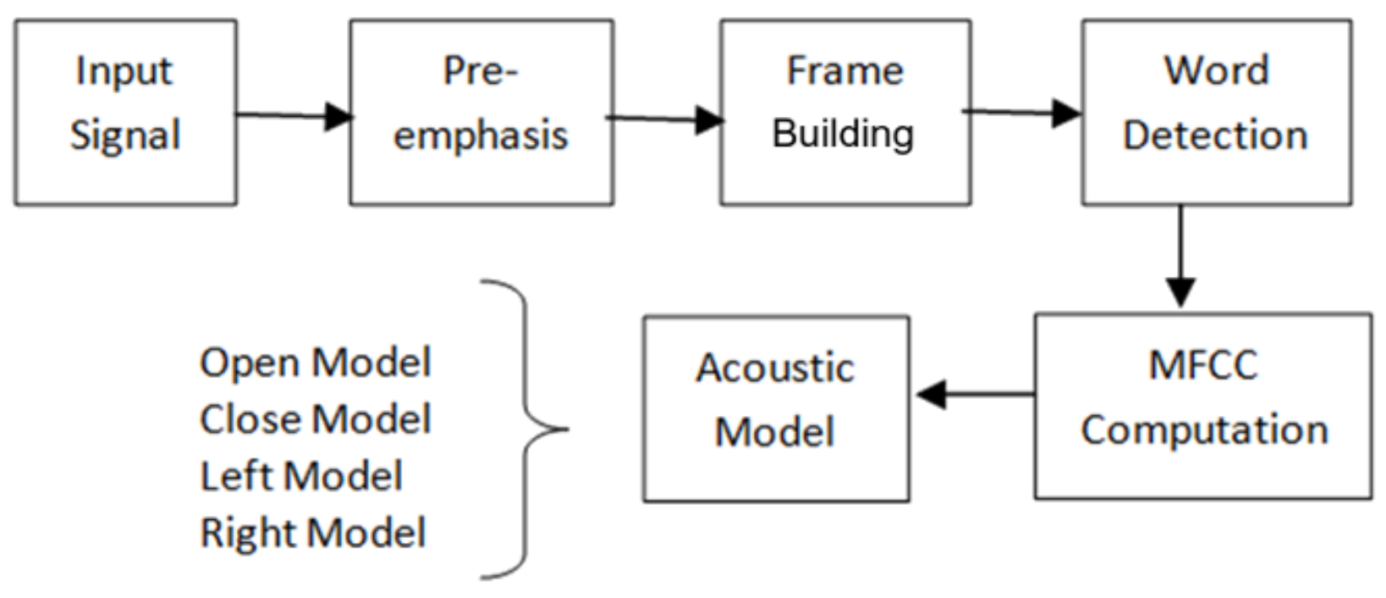

Figure 1 : Training Stage

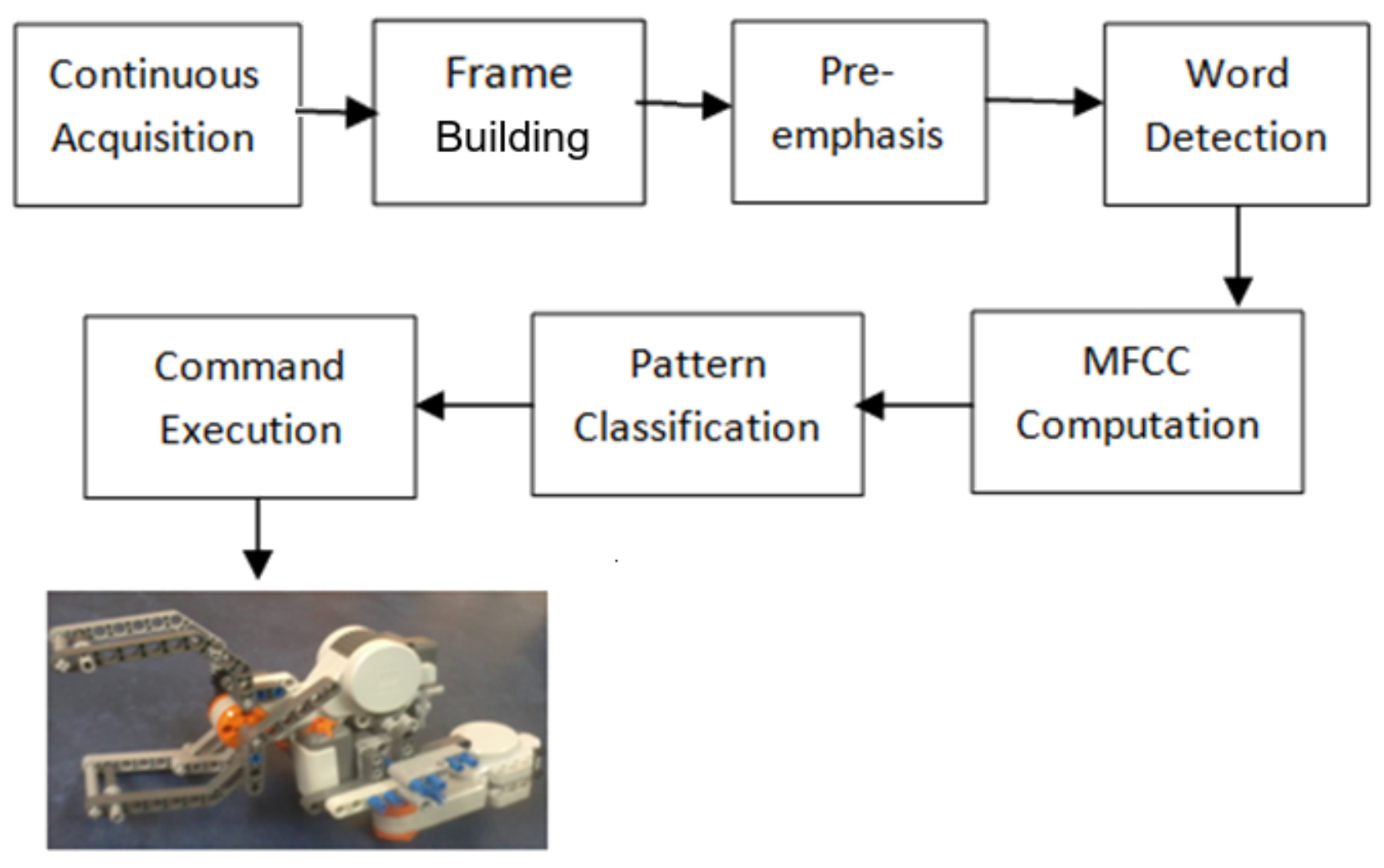

Figure 2 : Testing Stage.

In training stage, after applying a high-pass filter (Pre-emphasis), the signal is split in sets of 160 samples and an energy analysis is performed in order to detect intervals where a word exists. Then, Mel Frequency Cepstral Coefficients are computed and a space vector of MFCC are obtained, and the final step is to estimate a multidimensional probability density function, which will be the acoustic model for a word and a speaker. This model is generated based on Gaussian Mixture Models, using eight gaussians.

Testing stage performs the same steps described for training stage, from pre-emphasis to MFCC computation. The calculated coefficients vector is used as input to pattern classification step. The output parameters of this step are the posterior probability and the negative likelihood of the data given the model. This process is based on equations 1 and 2, where v is the coefficients vector, λ is the model, w_i are weights and N corresponds to normal distribution with mean μ and variance $\sum$. M is the number of mixtures. The *posterior* function of Matlab estimates the index $\hat{J}$ of the model that maximizes the conditional probability.

$$p(v|\lambda) = \sum_{i=1}^{M} w_i N(v|\mu_i, \sum_i) \tag{1}$$

$$\hat{J} = max_j p(v|\lambda_j) \tag{2}$$

Finally, if there is a successfull detection of a word matching a model, a command is sent to the gripper.

Results

Table 1 shows the performance of the system for two speakers and for changes in the number of Gaussians. The analysis was developed with regard to each command (command performance) and finally with regard to all the commands (global performance).

Speaker	GMM Number	No. of words	Command Performance (%)				Global Performance (%)
			Open	Close	Left	Right	
1	5	17	94.11	100	94.11	100	**97.05**
1	8	17	100	88.23	82.35	94.11	**88.23**
1	11	17	94.11	82.35	82.35	94.11	**88.23**
2	5	18	100	94.11	94.11	100	**97.05**
2	8	16	100	100	100	100	**100**
2	11	18	100	100	100	100	**100**

Table 1 : Performance of the Voice Recognition System

In table 2 different approaches of speech recognition systems for voice commands are compared. The accuracy is over 85% in all the solutions. Our system, with 95.09%, is above the mean of the compared implementations.

Approach	Description	Performance
Tezer [6]	Used LPC and DTW. Implemented in MATLAB.	86%
Beritelli [3]	Used VQ. Proposed to noise-robust application	92%
Bedoya [2]	Used wavelts, HMMs and MFCCs	98%
Phokharatkul [5]	Used filter banks and Mel scale analysis	96.3%
Ali [1]	Used MFCCs and ANN. Isolated or continuous speech mode	96%
Chin [4]	Used MFCCs and ANN	98.9%
Ours	Used MFFCs and GMMs, implemented in MATLAB	95.09%

Table 2 : Comparision of different speech recognition systems

Conclusions

In this paper we presented an speaker-dependent speech recognition system based on Mel Frequency Cepstral Coefficients (MFCCs) for extracting features and Gaussian Mixture Models (GMMs) for creating the model of each command. We test the systems with two different speakers and the worst case was 91.17% (average of global performance for speaker 1) of accuracy for a speaker. For a particular command, the worst case was 82.35% and the average of the global performance of the system was 95.09% (see Results section).

As future work, we propose the evaluation of the system with more than four commands and the creation of models with utterances from different speakers in order to test the capability of the system to be speaker-independent.

References

[1] S. Ali, S. Iqbal, and I. Saeed. Voice controlled urdu interface using isolated and continuous speech recognizer. In *Multitopic Conference (INMIC), 2012 15th International*, pages 53–57, 2012.

[2] W. Bedoya and L. Munoz. Methodology for voice commands recognition using stochastic classifiers. In *Image, Signal Processing, and Artificial Vision (STSIVA), 2012 XVII Symposium of*, pages 66–71, 2012.

[3] F. Beritelli and S. Serrano. A robust low-complexity algorithm for voice command recognition in adverse acoustic environments. In *Signal Processing, 2006 8th International Conference on*, volume 3, pages –, 2006.

[4] C. K. On, P. Pandiyan, S. Yaacob, and A. Saudi. Mel-frequency cepstral coefficient analysis in speech recognition. In *Computing Informatics, 2006. ICOCI '06. International Conference on*, pages 1–5, 2006.

[5] P. Phokharatkul, K. Nantanitikorn, and S. Phaiboon. Thai speech recognition using double filter banks for basic voice commanding. In *Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference on*, volume 6, pages 33–36, 2010.

[6] H. Tezer and M. Yagimli. Navigation autopilot with real time voice command recognition system. In *Signal Processing and Communications Applications Conference (SIU), 2013 21st*, pages 1–4, 2013.

Contact:
1. velasco.gustavo@correounivalle.edu.co
2. andres.a.diaz@correounivalle.edu.co

Perception and Intelligent Systems Research Group
School of Electric and Electronics Engineering
Universidad del Valle