Scientific Research Publishing

Journal of Computer and Communications, 2020, 8, 28-34
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Arabic Speech Recognition System Based on MFCC and HMMs

Hussien A. Elharati[1], Mohamed Alshaari[2], Veton Z. Këpuska[2]

[1]Electrical Engineering Department, High Institute of Science and Technology, Sūqal-Jum'a, Tripoli, Libya
[2]Electrical & Computer Engineering Department, Florida Institute of Technology, Melbourne, FL, USA
Email: hussien.elharati@gmail.com, malshaari2016@my.fit.edu, vkepuska@fit.edu

How to cite this paper: Elharati, H.A., Alshaari, M. and Këpuska, V.Z. (2020) Arabic Speech Recognition System Based on MFCC and HMMs. *Journal of Computer and Communications*, **8**, 28-34.
https://doi.org/10.4236/jcc.2020.83003

Received: January 23, 2020
Accepted: March 2, 2020
Published: March 5, 2020

Abstract

Speech recognition allows the machine to turn the speech signal into text through identification and understanding process. Extract the features, predict the maximum likelihood, and generate the models of the input speech signal are considered the most important steps to configure the Automatic Speech Recognition System (ASR). In this paper, an automatic Arabic speech recognition system was established using MATLAB and 24 Arabic words Consonant-Vowel Consonant-Vowel Consonant-Vowel (CVCVCV) was recorded from 19 Arabic native speakers, each speaker uttering the same word 3 times (total 1368 words). In order to test the system, 39-features were extracted by partitioning the speech signal into frames ~ 0.25 sec shifted by 0.10 sec. in back-end, the statistical models were generated by separated the features into number of states between 4 to 10, each state has 8-gaussian distributions. The data has 48 k sample rate and 32-bit depth and saved separately in a wave file format. The system was trained in phonetically rich and balanced Arabic speech words list (10 speakers * 3 times * 24 words, total 720 words) and tested using another word list (24 words * 9 speakers * 3 times *, total 648 words). Using different speakers similar words, the system obtained a very good word recognition accuracy results of 92.92% and a Word Error Rate (WER) of 7.08%.

Keywords

Speech Recognition, Feature Extraction, Maximum Likelihood, Gaussian Distribution, Consonant-Vowel

1. Introduction

Speech is a way to express ourselves, it's a complex naturally acquired human

motor ability [1]. Speech recognition is the capability of a device to receive, identify, and recognize the speech signal [2]. Speech recognition process fundamentally functions as a pipeline that converts the sound into recognized text, as shown in **Figure 1**. Based on spectral, the input signal is converted into a sequence of training and testing feature vectors saved in unique files. Given all the observations in the training data, Baum-Welch algorithm can learn and generate the HMM models equal to the number of the words to be recognized. In testing process, pattern matching provides likelihoods of a match of all sequences of speech recognition units to the input speech. Decision making generated according to the best path sequence between the models and testing data. Speech recognition system involved in several applications such as: call routing, automatic transcriptions, information searching, data entry, Speech to Text conversion, Text to Speech conversion etc. [3].

Arabic is the native language for over 300 million speakers and considered one of the official languages in many countries around the world. It has a unique set of diacritics that can change the meaning [4]. Arabic ASR received little attention compared to other languages, and research was oblivious to the diacritics in most cases. Omitting diacritics circumscribes the Arabic ASR system's usability for several applications such as voice-enabled translation, text to speech, and speech-to-speech [5].

Feature Extraction is accomplished by changing the waveform speech form to a form of parametric representation with a relatively low data rate for subsequent processing and analysis. Subsequently, the acceptable classification in the training and testing part is derived from the quality features [6]. Therefore, the most popular speech methods, Mel Cepstral frequency coefficients (MFCC) and Hidden Markov Model have been selected and tested in order to provide a high level of reliability and acceptability of the Arabic ASR.

2. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a feature widely used in automatic speech and speaker recognition has

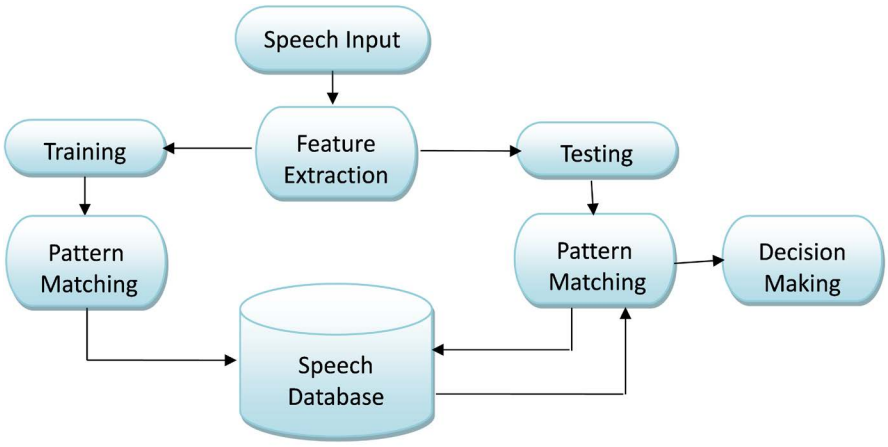

Figure 1. Speech recognition process.

been used to extract spectral features from frame sequences [7] [8]. Fast Fourier Transform (FFT) has been used to transfer the signal into frequency domain using the Equation (2.1). After pre-emphases, blocking, and windowing the input signal, FFT applies on the speech frames to obtain 256-point certain parameters, converting the power-spectrum to a Mel-frequency spectrum using Equations (2.2) and (2.3), and finally taking the logarithm of that spectrum and computing its inverse Fourier transform as shown in **Figure 2**.

$$X(k) = \sum_{N=0}^{N-1} x[n] e^{-j2\pi nk/N}, \quad 0 \le k < N \tag{2.1}$$

$$F_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{2.2}$$

$$F_{Hz} = 700 \cdot \left(10^{\frac{F_{mel}}{2595}} - 1\right). \tag{2.3}$$

3. Hidden Markov Model (HMM)

HMM is used to classify the features and generate the correct decision. HMM considered the powerful statistical tool used in speech recognition and speaker identification systems, due to the ability to model non-linearly aligning speech and estimating the model parameters [9]. Gaussian Mixtures also used to model the emission probability distribution function inside each state.

In training process, the observation parameters, transition probability matrix, the prior probabilities, and Gaussian distribution were re-estimated in order to get good parameters at each iteration as shown in **Figure 3**. As a result, all the previous HMM parameters are used to generate the likelihood scores, which are used to find the best path between the frames in order to recognize the unknown word [10] [11].

3.1. Evaluation Process

Given the observation sequence (O) and the model parameters (λ), Forward (α) and Backward (β) algorithms were used to find the probability of the observation sequence given the model $P(O|\lambda)$ [12]. As shown in **Figure 4**, forward and backward probabilities are added to evaluate the probability that any sequence of states has produced the sequence of observations.

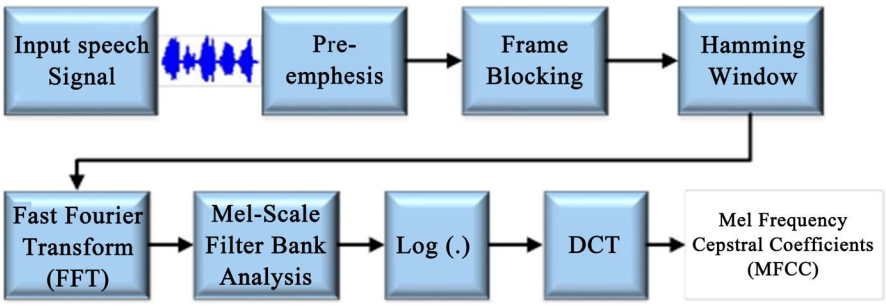

Figure 2. Mel Frequency Cepstral Coefficients (MFCC) block diagram.

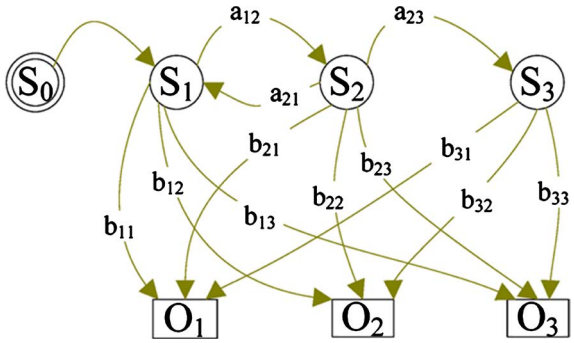

Figure 3. Three states hidden Markov model.

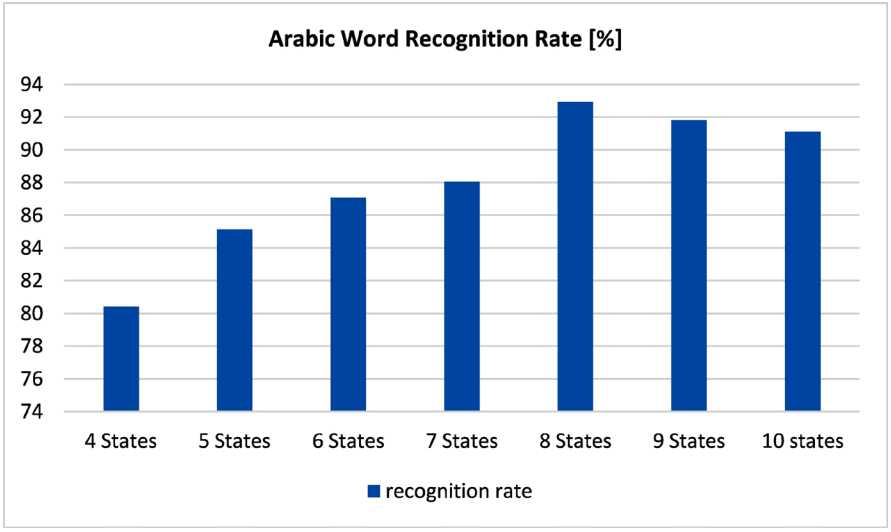

Figure 4. Recognition rate using different state numbers based on MFCC.

3.2. Training Process

Given the observation sequence (O) and the model parameters (λ), Baum-Welch algorithm was used to re-adjust and re-estimate the transition probability matrix and Gaussian mixture parameters (mean and covariance) that best describe the process [13] [14]. Baum welch algorithm also used to learn and encode the characteristics of the observation sequence in order to recognize a similar observation sequence.

3.3. Decoding Process

Viterbi algorithm has been used to comparing between the training and the testing data and find the optimal scoring path of state sequence by selecting the high probabilities between the model and the testing data [15] [16]. The maximal probability of state sequences is defined using the Equation (3.1), and the optimal scoring path of state sequence selected is calculated using the following MATLAB function.

start_recognition ("testing_list.mat", dim).

$$\delta t(i) = \max\left(P\left(q(1), q(2), \cdots, q(t-1); o(1), o(2), \cdots, o(t) \mid \lambda\right)\right) \quad (3.1)$$

4. Experimental Results

Using the automatic ASR system, several experiments were carried out using 24 (CVCVCV) Arabic isolated words as shown in Table 1. The feature vectors have been extracted for each sound using MFCC algorithm and saved, and the statistical models were generated using Hidden Markov Model classifier to match the data. The performance evaluation of the Arabic ASR system was obtained by finding the maximum word recognition rate.

In this work, (24 words * 3 times) Arabic CVCVCV words, small vocabulary data set are recorded from 19 adult male speakers (total 1368) divided into training and testing files. Table 2 shows the confusion matrix of the average

Table 1. CVCVCV arabic words.

Number	Word	Number	Word	Number	Word	Number	Word
1	فَعَلَ	7	فَعِلَ	13	فَعُلَ	19	فُعِلَ
2	رَفَعَ	8	بَخِلَ	14	بَلُغَ	20	ذُكِرَ
3	ذَكَرَ	9	عَمِلَ	15	صَلُحَ	21	جُمِعَ
4	ذَهَبَ	10	حَفِظَ	16	سَهُلَ	22	خُلِقَ
5	شَرَحَ	11	سَمِعَ	17	كَبُرَ	23	كُتِبَ
6	كَتَبَ	12	فَرِحَ	18	كَرُمَ	24	خُسِرَ

Table 2. Recognition rate using different state numbers based on MFCC.

State No.	Wrong words											
	1	2	3	4	5	6	7	8	9	10	11	12
4	3	4	12	5	0	8	27	8	2	2	0	3
5	6	1	11	4	0	6	1	0	0	2	0	1
6	4	3	4	0	0	5	0	0	4	0	1	5
7	12	0	4	2	1	3	4	3	2	0	2	2
8	3	0	1	1	1	3	0	0	1	0	0	0
9	0	1	2	1	0	1	3	1	0	0	0	1
10	8	0	2	1	0	0	0	0	0	2	0	1

State No.	Wrong words											
	13	14	15	16	17	18	19	20	21	22	23	24
4	5	0	7	3	0	10	3	6	0	0	11	22
5	17	0	10	3	2	0	16	8	0	0	3	16
6	4	1	5	10	5	0	21	7	0	1	1	12
7	11	0	5	9	3	0	6	5	0	0	0	12
8	10	0	4	4	0	1	9	2	0	0	1	10
9	9	0	8	4	2	0	4	8	0	0	1	13
10	13	2	4	3	0	0	13	0	0	1	7	7

Table 3. Recognition rate summary based on MFCC.

State No.	Total error count	Total correct count	Recognition rate
4	141	579	80.4166667
5	107	613	85.1388889
6	93	627	87.0833333
7	86	634	88.0555556
8	51	669	92.9166667
9	59	661	91.8055556
10	64	656	91.1111111

classification results, which obtained using convenient features in training and testing sessions. Each experiment conducted by dividing the data into 4, 5, 6, 7, 8, 9, and 10 number of states and modeled using 8 multi-dimensional Gaussians Hidden Markov Model.

During the experiments, the speech signal pre-reemphasis using 0.975 factor, covered by 25 milliseconds hamming window, and 10 milliseconds overlapping. The 256-point Fast Fourier Transform (FFT) was applied to the signal to transform 200 samples of speech from time to frequency domain. The summary of the resulting confidence level intervals for the recognition rate obtained in decoding process are listed in Table 3 and the chart in Figure 1 summarizes the recognition rate obtained for each state number.

5. Conclusion

The primary contribution of this work is to design Arabic ASR system and find the performance of the selected Arabic words is successfully verified and examined. For this purpose, 24 CVCVCV Arabic words were recorded from native speakers, all the experiments are conducted, and the recognition results of the ASR system were investigated and evaluated. The system is designed by MATLAB based on MFCC and discrete-observation multivariate HMM. In this work, the best results are achieved when the acoustic signals are extracted using 10 states and modeled by 8 Gaussian mixtures. The best recognition rate reaches 92.92% (51 total error count from 1368 total words count). According to Figure 3, the recognition rate decreased when using more or less than 10 state numbers.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Rabiner, L.R. and Juang, B.-H. (1993) Fundamentals of Speech Recognition. PTR Prentice Hall, Englewood Cliffs.

[2] Këpuska, V. and Klein, T. (2009) A Novel Wake-Up-Word Speech Recognition

System, Wake-Up-Word Recognition Task, Technology and Evaluation. *Nonlinear Analysis: Theory, Methods & Applications*, **71**, e2772-e2789. https://doi.org/10.1016/j.na.2009.06.089

[3] Këpuska, V.Z. and Elharati, H.A. (2015) Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model Classifier in Noisy Conditions. *Journal of Computer and Communications*, **3**, 1-9. https://doi.org/10.4236/jcc.2015.36001

[4] Satori, H., Harti, M. and Chenfour, N. (2007) Introduction to Arabic Speech Recognition Using CMUSphinx System. https://doi.org/10.1109/ISCIII.2007.367358

[5] Hyassat, H. and Zitar, R.A. (2006) Arabic Speech Recognition Using Sphinx Engine. *International Journal of Speech Technology*, **9**, 133-150. https://doi.org/10.1007/s10772-008-9009-1

[6] Këpuska, V.Z. and Elharati, H.A. (2015) Performance Evaluation of Conventional and Hybrid Feature Extractions Using Multivariate HMM Classifier. *International Journal of Engineering Research and Applications*, **5**, 96-101.

[7] Meng, Y. (2004) Speech Recognition on DSP: Algorithm Optimization and Performance Analysis. The Chinese University of Hong Kong, Hong Kong.

[8] Alkanhal, M.I., Al-Badrashiny, M.A., Alghamdi, M.M. and Al Qabbany, A.O. (2012) Automatic Stochastic Arabic Spelling Correction with Emphasis on Space Insertions and Deletions. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**, 2111-2122. https://doi.org/10.1109/TASL.2012.2197612

[9] Kumar, M., Aggarwal, R., Leekha, G. and Kumar, Y. (2012) Ensemble Feature Extraction Modules for Improved Hindi Speech Recognition System. *International Journal of Computer Science Issues*, **9**, 175.

[10] Abdelali, A., Darwish, K., Durrani, N. and Mubarak, H. (2016) Farasa: A Fast and Furious Segmenter for Arabic. In: *HLT-NAACL Demos*, Association for Computational Linguistics, San Diego, 11-16. https://doi.org/10.18653/v1/N16-3003

[11] Huang, X., Acero, A. and Hon, H.-W. (2001) Spoken Language Processing. Prentice Hall, Englewood Cliffs.

[12] Bogert, B.P., Healy, M.J. and Tukey, J.W. (1963) The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. *Proceedings of the Symposium on Time Series Analysis*, Vol. 15, 209-243.

[13] Sakoe, H. and Chiba, S. (1978) Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **26**, 43-49. https://doi.org/10.1109/TASSP.1978.1163055

[14] Dhingra, S.D., Nijhawan, G. and Pandit, P. (2013) Isolated Speech Recognition Using MFCC and DTW. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, **2**, 4085-4092.

[15] Okamoto, T., Hiroe, A. and Kawai, H. (2017) Reducing Latency for Language Identification Based on Large-Vocabulary Continuous Speech Recognition. *Acoustical Science and Technology*, **38**, 38-41. https://doi.org/10.1250/ast.38.38

[16] Ding, N., Melloni, L., Tian, X. and Poeppel, D. (2017) Rule-Based and Word-Level Statistics-Based Processing of Language: Insights from Neuroscience. *Language, Cognition and Neuroscience*, **32**, 570-575. https://doi.org/10.1080/23273798.2016.1215477