Efficient Feature Extraction Algorithms to Develop an Arabic Speech Recognition System

Abdulmalik A. Alasadi
Dept. of Computer Science and IT
Dr. Babasaheb Ambedkar Marathwada University
Aurangabad, India
dba.ora10g@gmail.com

Theyazn H. H. Adhyani
Community College in Abqaiq
King Faisal University
Saudi Arabia
taldhyani@kfu.edu.sa

Ratnadeep R. Deshmukh
Dept. of Computer Science and IT
Dr. Babasaheb Ambedkar Marathwada University
Aurangabad, India
rrdeshmukh.csit@bamu.ac.in

Ahmed H. Alahmadi
Department of Computer Science
Taibah University
Saudi Arabia
aahmadio@taibahu.edu.sa

Ali Saleh Alshebami
Community College in Abqaiq
King Faisal University
Saudi Arabia
aalshebami@kfu.edu.sa

Abstract—**This paper studies three feature extraction methods, Mel-Frequency Cepstral Coefficients (MFCC), Power-Normalized Cepstral Coefficients (PNCC), and Modified Group Delay Function (ModGDF) for the development of an Automated Speech Recognition System (ASR) in Arabic. The Support Vector Machine (SVM) algorithm processed the obtained features. These feature extraction algorithms extract speech or voice characteristics and process the group delay functionality calculated straight from the voice signal. These algorithms were deployed to extract audio forms from Arabic speakers. PNCC provided the best recognition results in Arabic speech in comparison with the other methods. Simulation results showed that PNCC and ModGDF were more accurate than MFCC in Arabic speech recognition.**

Keywords-speech recognition; feature extraction; PNCC; ModGDF; MFCC; Arabic speech recognition

I. INTRODUCTION

Speech is the most commonly and widely used form of communication. Many researches focus on developing reliable systems that can understand and accept commands through speech. Nowadays computers are involved in almost every aspect of our life, and as communication between people is mostly vocal, people anticipate the same way of interaction with computers [1]. Speech has the capacity to be an important mode of human-computer interaction, and the interest in developing computers that can accept speech as input is growing. The substantial research effort in global speech recognition and the increasing computational power at lower cost could result in more speech recognition applications in the near future [3]. Arabic language is the most popular in the Arab world, and the Arabic alphabet is used in some other languages such as Persian, Urdu, and Malaysian [2].

Research in human-computer speech interaction has focused mostly on developing better technical speech recognition systems, and gains in precision and productivity [4]. This research applied three distinct feature extraction methods onto an Arabic speech dataset, namely Mel-Frequency Cepstral (MFCC), Power-Normalized Cepstral Coefficients (PNCC) and Modified Group Delay Function (ModGDF). The extracted features were classified by a Support Vector Machine (SVM). The results of these three feature extracting techniques were compared in order to get the most efficient and accurate output. The feature extraction techniques, having their own properties like ModGDF, give additive and high-resolution signal. The additive property adds different functions in one group domain, and high-resolution property is used to sharpen the peaks of a group delay domain [5].

II. BACKGROUND

Speech awareness and evaluation have captivated researchers from Fletcher's early works [6] and the first voice identification devices [7], to present-day. Nevertheless, high precision machine speech recognition can be achieved mostly in quiet settings, as the efficiency of a typical speech recognizer reduces significantly in loud settings [8]. Environmental influence and other variables were explored in [9]. As technology progresses, speech recognition will be embedded in more devices used in everyday activities, where environmental variables perform a major part, such as mobile phone voice recognition applications [10], cars [11], integrated access control and information systems [12], emotion identification systems [13], application monitoring [14], disabled assistance [15], and intelligent technology. In addition to voice, many acoustic applications are also essential in diverse engineering issues [16–22]. A noise decrease method could be deployed to enhance efficiency in real-world noisy settings [23–26]. Machine efficiency degrades on noise,

Corresponding author: Theyazn H. H. Adhyani

channel variance, and spontaneous expressions further below than humans [27]. Automatic Speech Recognition (ASR) has not surpassed human efficiency in precision and robustness but we continue to avail from it by knowing the central values behind the identification of Human Speech (HS) [28]. Despite the advancements in auditory processing and popular front-ends for ASR devices, only a few elements of noise handling in the auditory periphery are modeled and simulated [29]. For instance, common methods such as MFCC use auditory features like varying bandwidth filter bank and compression size. Coefficients of Perceptual Linear Prediction (PLP) focus on perceptual processing by using curves of critical band resolution, corresponding loudness scaling, and cube root energy laws of listening Linear Prediction Coefficients (LPC) [30]. Synaptic adjustment could include an instance of auditory-motivated enhancements of voice depiction. Standard MFCC or PLP coefficients could be substituted by coefficients depending on some cochlear model in order to better represent human auditory periphery. The proposed model of synaptic adaptation in [31] showed important improvements in the efficiency of speech recognition. The PNCC proposed in [32], was based on auditory processing, including new characteristics, using a nonlinearity of power-law, a noise-suppression algorithm relying on asymmetric filtering, and temporal masking. The experimental findings exhibited enhanced precision of acceptance, comparing to MFCC and PLP. Another strategy for feature removal was based on Deep Neural Networks (DNN). The noise robustness of sound designs relying on DNN was evaluated in [33]. Recurrent Neural Networks (RNN) for cleaning distorted input characteristics were applied in [34]. The use of LSTM-RNNs was suggested in [35] to manage extremely non-stationary additive noise. For solid voice recognition, an all-inclusive outline of profound teaching was presented in [36]. Many researches utilized PNCC and MFCC to extract the most significant features from speech signals [37-39]. Group Delay Function (ModGDF) was used to extract speech signals, being more efficient than MFCC.

III. METHOD

Figure 1, shows the developed recognition system for evaluating the identification of Arabic speech.

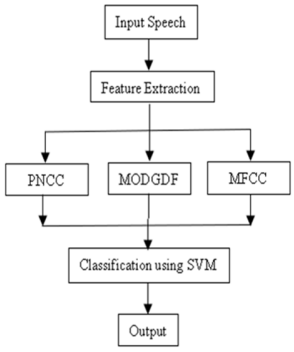

Fig. 1. Proposed speech recognition system

Audio from Arabic speakers was given as input to the system, and three feature extraction techniques, MFCC, PNCC and ModGDF, were applied to extract significant features of Arabic speech. SVM algorithm was used for training and classification, and performance measures were employed to evaluate these algorithms.

IV. DATABASE

A speech database was created, populated with utterances from volunteered Yemeni students studying at Dr. Babasaheb Ambedkar Marathwada University, in Aurangabad, India. Tables I and II, include the demographic information of the volunteers and the basic parameters of the recordings.

TABLE I. DEMOGRAPHICS OF VOLUNTEERS

Parameter	Values
Speaker type	Students (BSc, MSc, PhD)
Gender	35 Male, 15 Female
Basic language	Arabic
Accent	Standard and Yemeni
Age group	20 - 35
Country	Yemen
Environment	Dept. of CS & IT

TABLE II. BASIC RECORDING PARAMETERS

Parameter	Value
Sampling rate	16000Hz
Speakers	Dependent
Condition of noise	Normal
Accent	Arabic
Pre–emphasis	1-0.9/(z-1)
Window type	Hamming, 25ms
Window step size	20ms

A. Recording Procedure

The database was recorded using high quality headsets (Sennheiser PC360) and PRAAT Software, in a quiet environment. Speech samples were recorded in mono mode with 16000Hz sampling rate. A microphone was placed at a distance of about 3cm from the volunteer's mouth. Table III, displays the hardware and software used during the speech samples recording procedure.

TABLE III. HARDWARE AND SOFTWARE DETAILS

Hardware	Software
Laptop Hp Elite Book: (Core i7 ,5th gen, 8GB RAM, SSD 500GB)	Windows 10
Headphone :Sennheiser PC360	PRAAT: 6102_win64
Microphone	

B. Isolated Digits

Table IV shows the recorded Arabic digits.

C. Isolated Words

Isolated Arabic words of the speech corpus were used. Table V shows the Arabic words related to learning.

D. Continuous Sentences

Table VI shows the continuous sentence text corpus. Five utterances were collected for each sentence.

TABLE IV. ARABIC DIGITS

Digit	Pronunciation	Arabic writing
0	Safer	صفر
1	Wahed	وَاحِد
2	Ethnan	اِثْنَان
3	Thlathah	ثَلَاثَة
4	Arbaah	أَرْبَعَة
5	Khamsah	خَمْسَة
6	Settah	سِتَّة
7	Sabaah	سَبْعَة
8	Thamaneyah	ثَمَانِيَة
9	Tesaah	تِسْعَة

TABLE V. ARABIC WORDS

Arabic Word	Arabic pronunciation	English word
جامعة	Jameaah	University
كلية	Koleyah	Collage
قسم	Kesm	Department
تعليم	Taaleem	Education
محاضر	Mauhader	Lecture
مدرس	Modares	Teacher
معمل	Maamal	Lab
مادة	Madah	Course

TABLE VI. ARABIC SENTENCES RELATED TO GREETINGS

English language	Arabic language
When does registrations begin at the university?	متى يبدأ التسجيل في الجامعة ؟
Is there a graduate department?	هل يوجد قسم للدراسات العليا ؟
What are the admission requirements?	ما هي شروط القبول ؟
Is there a university website?	هل يوجد موقع الكتروني للجامعة؟
What are the available majors?	ما هي التخصصات المتوفرة ؟
The University has modern programs.	الجامعة لديها برامج حديثة
The mission of the university is ambitious.	رسالة الجامعة طموحة

V. FEATURE EXTRACTION ALGORITHMS

Feature extraction is vital for developing a speech recognition system. Its main objective is to extract the most significant features for identifying Arabic speakers. Three feature extraction algorithms were applied: PNCC, ModGDF, and MFCC.

A. Power Normalized Cepstral Cofficients (PNCC)

The PNCC feature extraction algorithm for extracting features for speech recognition can be seen in [3]. PNCC has two components: initial processing, and temporal integration for environmental analysis.

1) Initial Processing

This processing uses a pre-emphasis filter in the form of:

$$H(z) = 1 - 0.97z - 1 \quad (1)$$

Subsequently, a Short-Time Fourier Transformation (STFT) is conducted using Hamming windows. The use of a DFT volume of 1024 was intended to produce a length of 25.6ms, with 10ms between frames. By weighting magnitude-squared STFT outputs, spectral power in 40 analysis bands was obtained for positive frequencies. Center frequencies are also linearly spaced between 200Hz and 8000Hz using gamma tone filters in Equivalent Rectangular Bandwidth (ERB) [3].

2) Temporal Integration for Environmental Analysis

Most speech recognition systems use length frames of analysis between 20 and 30ms. It is often found that longer analytical windows deliver greater noise modeling efficiency and environmental normalization [6], because of the facility related to most background conditions, and changes slower than the speaking-related instant power. In PNCC processing, an estimate is made of a quantity referred to as "medium-time power" Q[m,l] by calculating the running average of P[m, l], the power observed in a single frame of analysis, according to:

$$\bar{Q}[m,l] = \frac{1}{2M+1} \sum_{m=m-M}^{m+M} P[m'l] \quad (2)$$

where m is the index of the frame, and l is the index of the channel.

B. Modified Group Delay Function (ModGDF)

This method was discussed in detail in [7-15]. It should be noted that the group delay feature is different from the phase spectrum, and it is defined as the phase negative derivative which can be used effectively to extract different system parameters when the signal is considered as a minimum phase signal. This is mainly so because a minimum phase signal's magnitude spectrum is similar to each other and its group delay feature. Figure 2, shows the process of ModGDF algorithm for extracting speech features. The algorithm is described below.

Algorithm: ModGDF feature extraction pseudocode
Input: speech x(n)
Output: ModGDF (Features vector) c(n)
Begin
Initialize parameters;
Apply the DFT of the speech x (n) as X[k];
Apply the DFT of the speech n x (n) as Y[k];
Calculate Group delay function where R and I represents real and imaginary parts;
Compute the spectrally smoothed spectra of X [k] and designate it as S [k];
Compute modified group delay where S [k] is the smoothed version of X [k] and two new parameters α and γ are used to regulate the dynamic range of ModGDF;
Apply the DCT to get the ModGD features;
Obtain ModGD Features vector (13 Coefficients for each frame);
End.

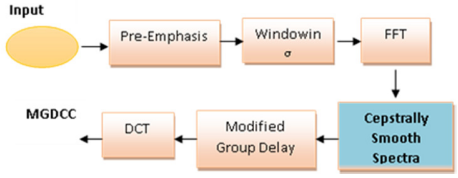

Fig. 2. Feature extraction process of ModGDF

C. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is the mostly used method in speech technology development, as it is similar to the human auditory system [16], taking into account its characteristics. Moreover, these

coefficients are robust and reliable to variations of speakers and recording conditions. Figure 3 shows the processing steps of MFCC for feature extraction.

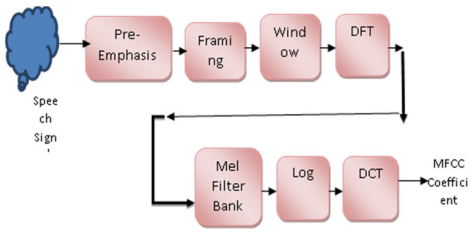

Fig. 3. Processes in MFCC feature extraction method

Pre-emphasis is the first step of MFCC, which produces energy, that was earlier compressed during sound generation, at a high frequency. Framing uses narrower parts to trim the sound signals. Windowing is used to avert discontinuity of the signals produced by the framing method. Fast Fourier Transform (FFT) is used for adapting a signal from time to frequency domain. Filter bank is the overlapping band pass filter. The final process is the Discrete Cosine Transform (DCT) making the coefficients of MFCC [18]. MFCC is computed from speech signal using the following three steps:

- Compute the FFT power spectrum of the speech signal
- Apply a Mel-space filter-bank to the power spectrum to get energies
- Compute DCT of log filter-bank energies to get uncorrelated MFCC's

The speech signal is first divided into time frames comprising of a random number of samples. In most systems, overlapping of frames is used to smooth transition from frame to frame. Each time frame is then windowed with a Hamming window to eliminate discontinuities at the edges [17]. The filter coefficients *w(n)* of a Hamming window of length *n* are computed according to:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

$$w(n) = 0, \text{ otherwise.} \quad (2)$$

where *N* is the total number of samples, and *n* is the current sample. Mel scale links perceived frequency or pitch of a pure tone to its actual measured frequency. Humans discern better small changes in pitch at lower frequencies. Integrating this scale makes the features match more closely to what humans hear. The formula for converting from frequency to Mel scale is:

$$M(f) = 1125\ln(1 + \frac{f}{700}) \quad (3)$$

while the formula to go back from Mel's scale to frequency is:

$$M^{-1}(m) = 700(\exp\left(\frac{m}{1125}\right) - 1) \quad (4)$$

VI. CLASSIFICATION

SVM is principally a binary classifier, but with the following two approaches it can be extended to multi-class tasks, the first being 1-vs-all i.e. comparing each class to the rest and the second, 1-vs-1, i.e. each class to the other, separately [20]. In this study, the i-vs-all was used, consisting of multiple binary SVMs equal to the number of classes. Every SVM with each one of the classes against the rest of them is taught and taken into consideration when testing. The decision is eventually made based on the distance from all SVMs between the test data and the hyper planes.

VII. SIMULATION RESULTS

Several experiments were conducted, employing the speech database, for classification and recognition using MFCC, PNCC and ModGDF for feature extraction. Training procedure used 60% of the data, while 40% were used for testing. The test procedure was implemented in Matlab 2016, and screen shots are shown in Figures 4 and 5. Evaluation and testing was performed using accuracy rate, specificity, sensitivity, precision, and execution time.

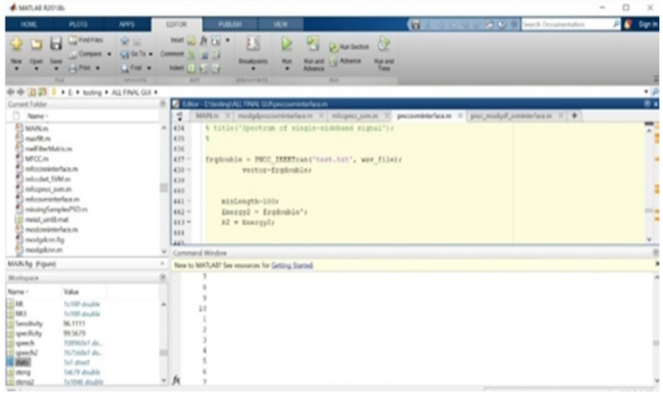

Fig. 4. Layout of the main system

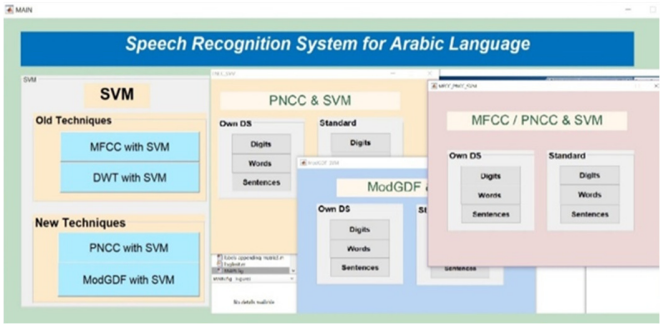

Fig. 5. Implementation

A. Analysis for Arabic Digits

The feature extraction methods were applied on the digit samples, and the results are shown in Table VII.

TABLE VII. SVM RESULTS ON DIGITS

Feature extraction technique	Accuracy rate	Specificity	Sensitivity	Precision	Execution time (s)
ModGDF	90.3	94.5	50.5	72.7	16.39
PNCC	97.5	98.6	87.6	88.7	54.8
MFCC	88.3	93.5	41.7	53.7	87.5

Figure 6 illustrates the methods' performance. As it can be observed, ModGDF with SVN obtained better results regarding time cost. PNCC and MFCC with SVN obtained good results, but their execution time was much higher. It is concluded that ModGDF had lower time cost, as it reduced execution time complexity. Table VIII, shows the confusion matrix of PNCC for the recognition of Arabic digits. Figure 7, displays a sample of ModGDF with SVM for the recognition of an Arabic digit ("Khamsah").

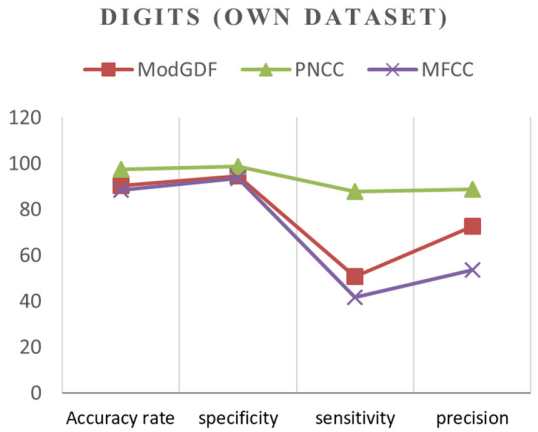

Fig. 6. Methods' performance on the recognition of Arabic digits

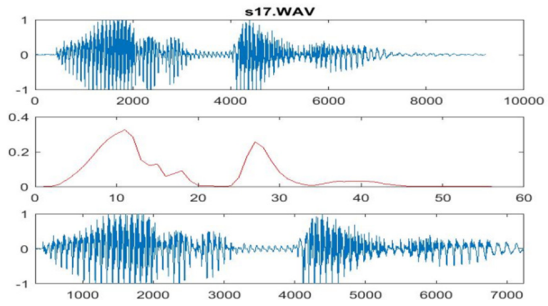

Fig. 7. ModGDF sample recognizing the Arabic digit "Khamsah"

TABLE VIII. CONFUSION MATRIX OF DIGITS USING PNCC/SVM

19	0	0	0	0	0	0	0	0	0
0	**19**	0	0	0	0	0	0	0	0
0	2	**17**	0	0	0	0	0	0	0
1	0	1	**17**	0	0	0	0	0	0
0	1	0	0	**18**	0	0	0	0	0
0	1	1	0	0	**17**	0	0	0	0
0	0	0	2	0	1	**16**	0	0	0
2	0	1	1	0	1	1	**13**	0	0
1	0	0	3	0	1	0	1	**13**	0
0	0	0	1	0	0	1	0	0	**17**

B. *Analysis for Arabic Words*

Table IX shows the results on the recognition of Arabic words. The results of ModGDF with SVM are reported to be not satisfactory, but time cost is much lesser than the other feature extraction methods. PNCC with SVM performed better, but time cost turned out to be significantly more. The results are also shown in Figure 8. Table X shows the confusion matrix of PNCC/SVM for the recognition of Arabic words. The confusion matrix has attested that PNCC is more robust and demonstrates more strength to identify Arabic words. Figure 9 illustrates the performance of the PNCC on the recognition of an Arabic word ("Dirham").

TABLE IX. RESULTS ON WORDS

Feature extraction technique	Accuracy rate	Specificity	Sensitivity	Precision	Execution time (s)
ModGDF	89.3	94.1	46.8	58.6	12.3
PNCC	95.15	97.3	75.8	79.2	49.5
MFCC	88.6	93.6	43.1	51.8	99.5

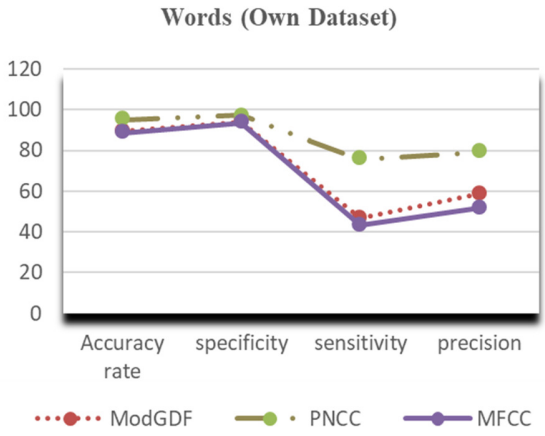

Fig. 8. Performance on the recognition of Arabic words

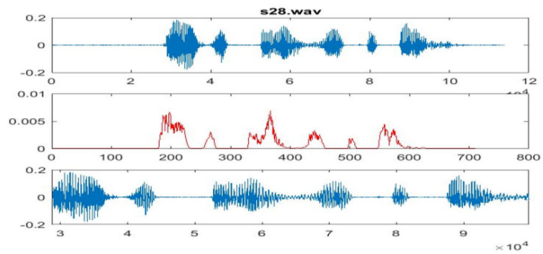

Fig. 9. Sample of PNCC with the SVM recognizing the Arabic word "Dirham"

TABLE X. CONFUSION MATRIX OF WORDS USING PNCC/SVM

19	0	0	0	0	0	0	0	0	0
2	**17**	0	0	0	0	0	0	0	0
2	2	**15**	0	0	0	0	0	0	0
2	1	1	**15**	0	0	0	0	0	0
2	0	3	2	**12**	0	0	0	0	0
1	1	2	2	0	**13**	0	0	0	0
1	3	1	0	0	0	**14**	0	0	0
0	0	0	1	2	0	0	**16**	0	0
0	1	1	1	0	0	0	0	**16**	0
2	0	0	0	0	0	2	2	1	**12**

C. *Analysis for Arabic Sentences*

Table XI shows the performance results on the recognition of Arabic sentences. As it can be observed, PNCC with SVM

performed better, but had greater execution time. PNCC had again the highest accuracy and lower execution time than MFCC. ModGDF had the lowest execution time of 12.3s, and accuracy of 88.2, while MFCC showed again the lowest accuracy. The accuracy of PNCC/SVM can be also confirmed by its confusion matrix analysis of sentences in Table XII. The confusion matrix shows that the PNCC/SVM is capable of recognizing sentences with satisfactory results. The results are also shown in Figure 10, while Figure 11 illustrates the performance of PNCC on the recognition of an Arabic sentence ("What are the available majors?").

TABLE XI. RESULTS ON SENTENCES

Feature extraction technique	Accuracy rate	Specificity	Sensitivity	Precision	Execution time (s)
ModGDF	88.2	93.5	41.2	45.3	18.9
PNCC	93.05	96.14	65.26	71.04	70.0
MFCC	86.0	92.2	30.0	49.48	125.0

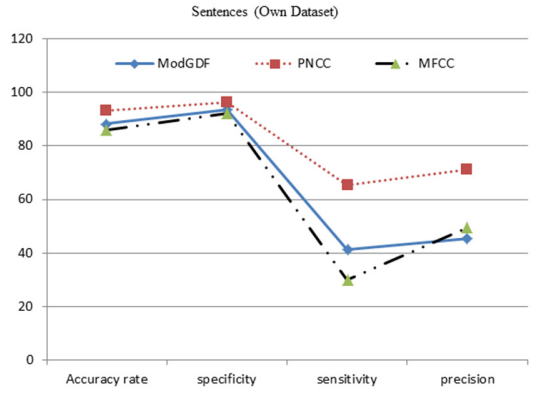

Fig. 10. Feature extraction performance on Arabic sentences

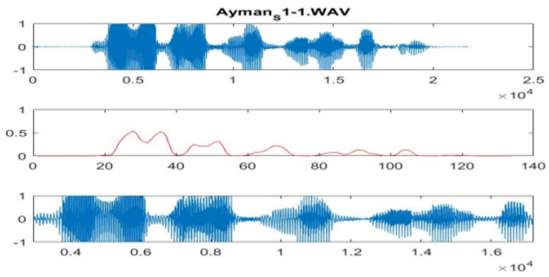

Fig. 11. Sample of PNCC/SVM recognizing the Arabic sentence "What are the available majors?"

TABLE XII. CONFUSION MATRIX OF SENTENCES USING PNCC/SVM

19	0	0	0	0	0	0	0	0	0
2	**17**	0	0	0	0	0	0	0	0
1	7	**11**	0	0	0	0	0	0	0
0	0	1	**18**	0	0	0	0	0	0
1	0	0	8	**10**	0	0	0	0	0
0	0	0	0	3	**16**	0	0	0	0
0	0	0	0	0	7	**12**	0	0	0
0	0	0	1	0	0	1	**17**	0	0
0	0	0	0	0	0	0	9	**10**	0
6	0	0	0	0	0	0	0	3	**10**

VIII. CONCLUSION

In this paper, a speech recognition system for Arabic language was presented, evaluating three feature extraction algorithms, namely MFCC, PNCC, and ModGDF, while an SVM was used for the classification process. Results showed that PNCC was more efficient, while ModGDF had moderate accuracy. PNCC and ModGDF fill the gaps in SVM, as they both had greater accuracy than MFCC. PNCC had a 93-97% accuracy rate, ModGDF had 90% and MFCC had 88%.

REFERENCES

[1] P. P. Shrishrimal, R. R. Deshmukh, V. B. Waghmare, "Indian language speech database: A review", International Journal of Computer Applications, Vol. 47, No. 5, pp. 17-21, 2012

[2] S. K. Gaikwad, B. W. Gawali, P. Yannawar, "A review on speech recognition technique", International Journal of Computer Applications, Vol. 10, No. 3, pp. 16-24, 2010

[3] C. Huang, T. Chen, E. Chang, "Accent issues in large vocabulary continuous speech recognition", International Journal of Speech Technology, Vol. 7, No. 2-3, pp. 141-153, 2004

[4] M. A. Anasuya, S. K. Katti, "Speech recognition by machine: A review", International Journal of Computer Science and Information Security, Vol. 6, No. 3, pp. 181-205, 2009

[5] P. L. Garvin, P. Ladefoged, "Speaker identification and message identification in speech recognition", Phonetica, Vol. 9, No. 4, pp. 193-199, 1963

[6] G. Ceidaite, L. Telksnys, "Analysis of factors influencing accuracy of speech recognition", Elektronika ir Elektrotechnika, Vol. 105, No. 9, pp. 69-72, 2010

[7] Z. H. Tan, B. Lindberg, "Speech recognition on mobile devices", in: Mobile Multimedia Processing – WMMP 2008, Lecture Notes in Computer Science, Vol. 5960, Springer, 2010

[8] W. Li, K. Takeda, F. Itakura, "Robust in-car speech recognition based on nonlinear multiple regressions", EURASIP Journal on Advances in Signal Processing, 2007

[9] W. Ou, W. Gao, Z. Li, S. Zhang, Q. Wang, "Application of keywords speech recognition in agricultural voice system", Second International Conference on Computational Intelligence and Natural Computing, Wuhan, China, September 13-14, 2010

[10] L. Zhu, L. Chen, D. Zhao, J. Zhou, W. Zhang, "Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN", Sensors, Vol. 17, No. 7, 2017

[11] J. E. Noriega-Linares, J. M. Navarro Ruiz, "On the application of the raspberry Pi as an advanced acoustic sensor network for noise monitoring", Electronics, Vol. 5, No. 4, 2016

[12] M. Al-Rousan, K. Assaleh, "A wavelet-and neural network-based voice system for a smart wheelchair control", Journal of the Franklin Institute, Vol. 348, No. 1, pp. 90-100, 2011

[13] I. V. McLoughlin, H. R. Sharifzadeh, "Speech recognition for smart homes", in: Speech Recognition, Technologies and Applications, Intech, 2008

[14] A. Glowacz, "Diagnostics of rotor damages of three-phase induction motors using acoustic signals and SMOFS-20-EXPANDED", Archives of Acoustics, Vol. 41, No. 3, pp. 507-515, 2016

[15] A. Glowacz, "Fault diagnosis of single-phase induction motor based on acoustic signals", Mechanical Systems and Signal Processing, Vol. 117, pp. 65-80, 2019

[16] M. Kunicki, A. Cichon, "Application of a phase resolved partial discharge pattern analysis for acoustic emission method in high voltage insulation systems diagnostics", Archives of Acoustics, Vol. 43, No. 2, pp. 235-243, 2018

[17] D. Mika, J. Jozwik, "Advanced time-frequency representation in voice signal analysis", Advances in Science and Technology Research Journal, Vol. 12, No. 1, pp. 251-259, 2018

[18] L. Zou, Y. Guo, H. Liu, L. Zhang, T. Zhao, "A method of abnormal states detection based on adaptive extraction of transformer vibro-acoustic signals", Energies, Vol. 10, No. 12, 2017

[19] H. Yang, G. Wen, Q. Hu, Y. Li, L. Dai, "Experimental investigation on influence factors of acoustic emission activity in coal failure process", Energies, Vol. 11, No. 6, Article ID 1414, 2018

[20] L. Mokhtarpour, H. Hassanpour, "A self-tuning hybrid active noise control system", Journal of the Franklin Institute, Vol. 349, No. 5, pp. 1904-1914, 2012

[21] S. C. Lee, J. F. Wang, M. H. Chen, "Threshold-based noise detection and reduction for automatic speech recognition system in human-robot interactions", Sensors, Vol. 18, No. 7, Article ID 2068, 2018

[22] S. M. Kuo, W. M. Peng, "Principle and applications of asymmetric crosstalk-resistant adaptive noise canceler", Journal of the Franklin Institute, Vol. 337, No. 1, pp. 57-71, 2000

[23] J. W. Hung, J. S. Lin, P. J. Wu, "Employing robust principal component analysis for noise-robust speech feature extraction in automatic speech recognition with the structure of a deep neural network", Applied System Innovation, Vol. 1, No. 3, Article ID 28, 2018

[24] R. P. Lippmann, "Speech recognition by machines and humans", Speech Communication, Vol. 22, No. 1, pp. 1-15, 1997

[25] J. B. Allen, "How do humans process and recognize speech?", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 567-577, 1994

[26] S. Haque, R. Togneri, A. Zaknich, "Perceptual features for automatic speech recognition in noisy environments", Speech Communication, Vol. 51, No. 1, pp. 58-75, 2009

[27] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", The Journal of the Acoustical Society of America, Vol. 87, No. 4, pp. 1738-1752, 1990

[28] M. Holmberg, D. Gelbart, W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 1, pp. 43-49, 2005

[29] C. Kim, R. M. Stern, "Power-normalized Cepstral Coefficients (PNCC) for robust speech recognition", 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan, March 25-30, 2012

[30] M. L. Seltzer, D. Yu, Y. Wang, "An investigation of deep neural networks for noise robust speech recognition", 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, May 26-31, 2013

[31] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR", 13th Annual Conference of the International Speech Communication Association, Portland, USA, September 9-13, 2012

[32] M. Wollmer, B. Schuller, F. Eyben, G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening", IEEE Journal of Selected Topics in Signal Processing, Vol. 4, No. 5, pp. 867-881, 2010

[33] Z. Zhang, J. Geiger, J. Pohjalainen, A. E. D. Mousa, W. Jin, B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments", ACM Transactions on Intelligent Systems and Technology, Vol. 9, No. 5, pp. 1-28, 2018

[34] E. Principi, S. Squartini, F. Piazza, "Power normalized cepstral coefficients based supervectors and i-vectors for small vocabulary speech recognition", 2014 International Joint Conference on Neural Networks, Beijing, China, July 6-11, 2014

[35] E. Loweimi, S. M. Ahadi, "A New group delay-based feature for robust speech recognition", 2011 IEEE International Conference on Multimedia and Expo, Barcelona, Spain, July 11-15, 2011

[36] B. Kurian, K. T. Shanavaz, N. G. Kurup, "PNCC based speech enhancement and its performance evaluation using SNR Loss", 2017 International Conference on Networks & Advances in Computational Technologies, Thiruvanthapuram, India, July 20-22, 2017

[37] T. Fux, D. Jouvet, "Evaluation of PNCC and extended spectral subtraction methods for robust speech recognition", 23rd European Signal Processing Conference, Nice, France, August 31 – September 4, 2015

[38] A. Kaur, A. Singh, "Power-Normalized Cepstral Coefficients (PNCC) for Punjabi automatic speech recognition using phone based modelling in HTK", 2nd International Conference on Applied and Theoretical Computing and Communication Technology, Bangalore, India, July 21-23, 2016

[39] C. Kim, R. M. Stern, "Feature extraction for robust speec recognition based on Mmximizing the sharpness of the power distribution and on power flooring", 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, USA, March 14-19, 2010

[40] D. S. Kim, S. Y. Lee, R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments", IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 1, pp. 55-69, 1999