TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 4, August 2020, pp. 2193~2204
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i4.14215

Arabic digits speech recognition and speaker identification in noisy environment using a hybrid model of VQ and GMM

Abdelkbir Ouisaadane[1], Said Safi[2], Miloud Frikel[3]
[1, 2]Department of Mathematics and Computer Science, Polydisciplinary Faculty,
Sultan Moulay Slimane University, Morocco
[3]ENSICAEN School, LAC Laboratory Caen-Normandie University, France

Article Info

Article history:

Received Sep 26, 2019
Revised Feb 24, 2020
Accepted Apr 20, 2020

Keywords:

Arabic digits
GMM
MFCC
SNR
VQ

ABSTRACT

This paper presents an automatic speaker identification and speech recognition for Arabic digits in noisy environment. In this work, the proposed system is able to identify the speaker after saving his voice in the database and adding noise. The mel frequency cepstral coefficients (MFCC) is the best approach used in building a program in the Matlab platform; also, the quantization is used for generating the codebooks. The Gaussian mixture modelling (GMM) algorithms are used to generate template, feature-matching purpose. In this paper, we have proposed a system based on MFCC-GMM and MFCC-VQ approaches on the one hand and by using the hybrid approach MFCC-VQ-GMM on the other hand for speaker modeling. The white Gaussian noise is added to the clean speech at several signal-to-noise ratio (SNR) levels to test the system in a noisy environment. The proposed system gives good results in recognition rate.

Corresponding Author:

Abdelkbir Ouisaadane,
Department of Mathematics and Computer Science,
Polydisciplinary Faculty,
Sultan Moulay Slimane University, Morocco.
Email: Abdelkbir.wiss@gmail.com

1. INTRODUCTION

Automatic speech recognition (ASR) is an important topic of speech processing. ASR is a technology that allows an electronic platform such as a smartphone or a computer to identify spoken words by humans [1]. Speech is a powerful and natural tool for communication. For this, the speech recognition system makes the interaction between a human and a machine more fluid and simpler [2]. In recent years, the researchers have developed more important research in biometric security technology with speaker recognition to make the communication between humans and machines to be more natural [3]. The speaker recognition system can be classified into identification and verification (recognition). Speaker identification is the process of automatically recognizing who is speaking based on individual information included in speech waves. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. This technique makes the speaker verify their identity and control such as security control, telephone shopping, access services to the voice mail, database access services and remote access to computers [4].

An ASR system involves two phases: the training phase and the testing phase. At the training phase, the parameters of the classification model are estimated using a large number of training data. The extraction of features is done from all speech signals using various feature extraction techniques such as MFCC, LPC, LDA, RASTA, etc [1, 3, 5]. These features are in the form of vectors is stored in reference models, in particular,

the acoustic model, which is used to characterize that word using the classification algorithm in the testing phase [5].

The most technique used for speech features comprises the Mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) and the linear predictive coding (LPC) coefficients. [6]. The MFCCs are the best known, for that reason we use it in this paper. The MFCCs are the best known, they are less susceptible to speaker-dependent variations and therefore we use it in this study. [7]. Many matching techniques are used in speech and speaker recognition, such as dynamic time warping (DTW), hidden Markov models (HMM) that are very frequently used in speech recognition [8, 9], artificial neural network (ANN), gaussian mixture model (GMM) and vector quantization (VQ) are generative models used for creating a speaker model. The Gaussian mixture model is widely used in speaker identification modeling system [10]. In this paper, the VQ method is employed and it will be compared with the GMM model. The employed method has characterized by its easiest implementation and its highest accuracy. For vector quantization (VQ) the LBG (Linde, Buzo, and Gray) algorithm and the k-means algorithm are the most familiar algorithms [11-13].

In voice applications, speech is damaged due to interference with background noise. Consequently, we cannot know whether the signal contains valid information or not through direct observation [14]. In this paper, the performance of a speaker identification system presented for the clean speech has been further investigated here by adding noise in particular additive white Gaussian noise (AWGN) to the clean 'speakers' utterances in the training and testing phase [15, 16].

Much works is done in speech processing for many languages in speaker system independent, automatic speech recognition for an isolated word or continuous speech such as [17-19], etc. Existing speech recognition systems are working well for European languages like English [20-24]. The researches for the Arabic language speech recognition is still weak, especially the continuous speech recognition in a noisy environment [25-29]. However, all applications in speech recognition are mostly available in English, like the works presented in [30-34]. Despite Arabic being the fourth most widely spoken language of the world, which is why our research was focused on Arabic speech recognition in noisy environments and because it is the first work using the methods that we mentioned earlier.

The Arabic language is the fourth largest language spoken by nearly 1.6 billion Muslims native speakers, this language spoken by the majority of the people in the Middle East and North Africa; note that Arabic has many different dialects. This is some little work in Arabic speaker and speech recognition [35-38]. We presented in Table 1 a Literature review for speech recognition research using MFCC, GMM and VQ techniques regarding Arabic or other languages.

- Background: speech recognition with speaker identification systems have widely extensive applied fields. Many works had performed in this area using multiple techniques. MFCC, HMM, GMM and VQ are the most prominent methods.
- Objectives: The aim of this paper is to execute a small-scale Arabic digit's speech recognition system in a noisy environment based on MFCC in features extraction and hybrid GMM-VQ for features training and classification. This system can recognize and respond to digits' speech inputs and compare an unknown speaker's speech against a database of N known speakers. The best match is returned as the identified speaker and the digit is spoken.
- The problem: The Arabic language is among the most spoken languages in the world with around 300 million native speakers. However, compared to other languages, the research is still poor in Arabic speech recognition, especially in a noisy environment. The presence of background noise, as well as the diversity of Arabic dialects, are considered challenges for Arabic speech recognition. Studying Moroccan Arabic, which is very difficult, which is even challenging in the orthographic rules, the multiple accents and vocabulary according to the regions of Morocco.
- The proposed solution: To contribute to developing Arabic speech recognition systems, we built two systems in one. The first for speaker identification and the second for Arabic spoken digits with AWGN background noise. Wherefore, we propose the hybrid GMM-VQ model along with MFCC as a feature extraction technique. Here VQ was used for training data and for speaker identification then the GMM for recognition. The efficacy of the proposed method is observed while performing different experiments and compared to earlier work.

The database for this work has been built using 100 male and female speakers, the vocabulary consists of 10 words representing the Arabic spoken digits from 0 to 9. MFCC technique has been used to extract the features and GMM, VQ, GMM-VQ models have been used for recognition. The system is tested by using test data spoken by 15 speakers and achieves an overall word-accuracy of 98.33% in clean condition using GMM+VQ.

The rest of this paper is organized as follows: In section 2, we clarify the system architecture in more depth. The experimental results are presented in section 3 followed by discussion in section 4. Finally, we indicate the conclusion and future work in section 5.

Table 1. Summary table of limited literature review on speech recognition studies done in Arabic and other languages

Author	Language	Year	Feature extraction	Method	Recogntion rate (%)
Giorgio Biagetti et al. [4]	English Timit	2017	Karhunen-Loève transform (DKLT)	EM, GMM	97.70% (noisy conditions)
Mohit Dua et al. [5]	Hindi	2018	MFCC, GFCC, BFCC	HMM-GMM	MFCC 65.25 %, GFCC 75.02%, BFCC 75.56 %
Bhadragiri Jagan Mohan and Ramesh Babu N. [7]	English	2014	MFCC	DTW	satisfying
T. K. Das et al. [8]	English	2016	MFCC	VQ, HMM	90%
D. Nagajyothi and P. Siddaiah [11]	Speaker voice English	2017	MFCC	VQ, LBG	high accuracy
Arnav Gupta and Harshit Gupta [12]	Speaker voice English	2013	MFCC	VQ	89%
Ankur Maurya et al. [13]	Hindi	2017	MFCC	VQ, GMM	85.49 % using MFCC -VQ 94.12 % using MFCC-GMM
Veena and Mathew [14]	English Timit	2015	MFCC	SVM-GMM	95% in clean, 90% in noisy
Musab Al-Kaltakchi et al. [15]	English Timit	2017	PNCC, MFCC	GMM-UBM	95% in clean, 75.83% SNR (0-30) dB
S. B. Dhonde and S. M. Jagade [17]	English Timit	2016	MFCC	VQ, GMM	98.4 % with MFCC –VQ 99. 2 % with MFCC-GMM
U. G. Patil et al. [18]	Hindi	2016	MFCC	VQ-GMM	94.31 %
S. Karpagavalli et al. [21]	Tamil	2012	MFCC	HMM	92%
Rafik Djemili et al. [22]	English IViE corpus	2012	MFCC	GMM, MLP, VQ, LVQ	96.4% with GMM, MLP, VQ 94.6% with LVQ
Bidhan Barai et al [23]	English	2017	MFCC, GFCC	VQ/GMM	100% in clean, 90% in noisy
Chen Wang et al [24]	English	2008	MFCC	VQ-GMM	93.1%.
N Hammami and M Bedda [25]	Arabic	2010	MFCC	VQ - MWST	93.12%
Awais Mahmood et al [26]	Arabic	2014	MFCC, MDLF	GMM	96.89%
M Alsulaiman et al [28]	Arabic	2016	MFCC, MDLF, and MDLF-MA	GMM	94%
Mohamed Khelifa et al [29]	Arabic	2017	MFCC	HMM /GMM	Between 94% and 97%
Azzedine Touazi and Mohamed Debyeche [35]	Arabic	2017	MFCC	HMM	99.89 % in clean, 95.94% in multi-condition
Anissa Imen Amrous et al [38]	Arabic	2011	MFCC	HMM	93.91 % in clean, 32.33% in Pink noise 5dB

2. THE SYSTEM ARCHITECTURE

2.1. Arabic digits speech recognition system

The first ten Arabic digits are : “Siffer”, “Wahed”, “Ithnani”, “Thalatha”, “Arbaa”, “Khamsa”, “Sitta”, “Sabaa”, “thamanya” and “tisaa”.The Arabic spoken digits would be helpful in many applications such as telephone dialing systems, banking systems, airline reservations, etc. These ten digits are polysyllabic words except “zero/siffer” which is a monosyllable word as shown in Table 2. The syllables in the Arabic language are CV, CVC, and CVCC information the decoder needs to do its job. V indicates a (long or short) vowel while C indicates a consonant. Arabic utterances can only start with a consonant [37].

Table 2. Arabic digits

Digits	Arabic writing	Pronunciation	Syllables	Number of syllables
0	صفر	sefr	CVCC	1
1	واحد	wa-hed	CV-CVC	2
2	اثنين	aath-nayn	CVC-CVC	2
3	ثلاثة	tha-la- thah	CV-CV-CVC	3
4	أربعة	aar-baah	CVC-CV-CVC	3
5	خمسة	kham-sah	CVC-CVC	2
6	ستة	set-tah	CVC-CVC	2
7	سبعة	sub-aah	CVC-CVC	2
8	ثمانية	tha-ma-nyeh	CV-CV-CVC	3
9	تسعة	tes-ah	CVC-CVC	2

2.2. The system architecture

Automatic speaker recognition system (ASR) was defined as the process of identifying a speaker by analyzing spectral shape of his voice signal. The Speaker recognition as illustrated in Figure 1 represents the process of identifying a person from his voice after recording it with a microphone and compares it with

another stored as training. This block system is split into two phases: the first one represents the training phase and the second one is testing. During these two phases, speaker identification consists of four steps: voice recording, feature extraction, pattern matching and decision (recognized/not recognized).

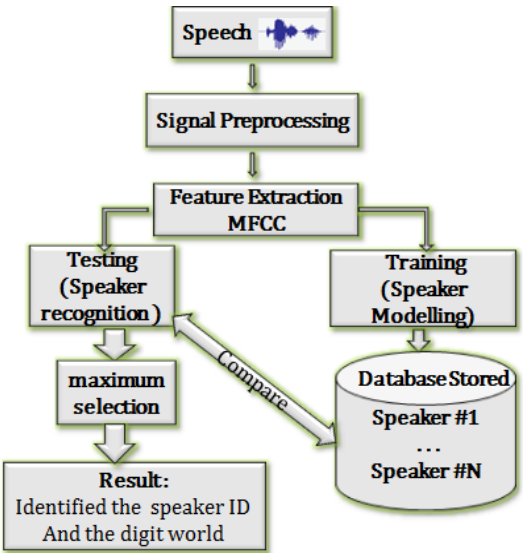

Figure 1. General structure of speech and speaker recognition system

The system is separated into two portions: speaker identification and spoken digit recognition. Where:

- Spoken digit recognition: to recognize the word among 10 Arab digits words.
- Speaker identification: to recognize the speaker for a particular spoken word.

2.2.1. Training phase

The speaker's reference database along with the speaker IDs and their audio recordings are stored; the system can build a reference model for that speaker. Regarding in training phase for spoken digit recognition, we previously recorded a database and converting it into acoustic vectors using Mel-frequency cepstrum coefficients (MFCC).

2.2.2. Testing phase

In the testing phase, the system checks that the speaker's input speech is similar at that which is stored in the reference. Therefore, the system can identify the person who is speaking and the digit that saying. During this phase, 450 voice samples are recorded by 10 male voices and 5 female voices chosen from our database. These clean test data is then mixed with white Gaussian noise (AWGN) with different levels of SNRs (5, 10, 15, 20 dB). The codebook vectors are developed using the proposed VQ-GMM approach from a specific speaker's voice. Then, they will be compared with the reference models obtained in the training phase.

2.3. Mel-frequency cepstrum coefficients (MFCC)

Mel-frequency cepstrum coefficients (MFCC) are popular features extracted from speech signals for use in recognition tasks. MFCCs are based on a perceptually scaled frequency axis. This also allows for better representation of the speech. The following relation is used to calculate the Mels scale for a given frequency f (Hz) of signal is given by (1).

$$Mel(f) = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

Figure 2 shows the general block diagram for extraction of MFCC features vectors. The basic five operations are carried on speech signal to get the cepstral coefficients.The acoustic features used in this evaluation are composed of 39 parameters with 13 MFCCs.

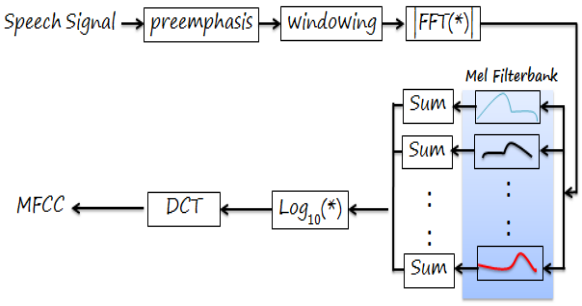

Figure 2. Detailed MFCC process

2.4. Vector quantization

Vector Quantization (VQ) is a classical and the most frequently used pattern-matching technique [2]. We will use the VQ approach, in this paper, due to its easiest implementation and its higher accuracy. This technique consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker-specific features. The training data features created by VQ method are combined to create a codebook for each speaker. In the recognition phase, the system compares the difference between a speaker's test data and the codebook of each speaker. Accordingly, it concludes the recognition result [39-41].

Figure 3 shows an illustrative diagram of this recognition process. One speaker can be discriminated against from another base of the location of centroids. In the training phase, using the clustering LBG algorithm [42]. In Figure 3 we are limited to present two speakers and her acoustic vectors. The yellow circles refer to the acoustic vectors from the speaker 1 while the blue circles are from the speaker 2. A speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result codewords (centroids) are shown in Figure 3 by black circles for speaker 1 and red circles for speaker 2. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. The VQ distortion illustrates the distance from the nearest codebook, calculated in the testing phase of speaker recognition system. The 'adequate' speaker corresponds to minimum VQ distortion, so it is selected and verified [42]. We used the LBG algorithm to build a codebook from a set of training vector for this purpose [42].

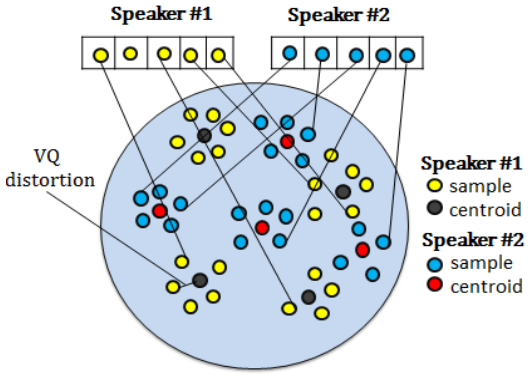

Figure 3. Vector quantization codebook

2.5. Gaussian mixture model

Gaussian mixture model (GMM) is one of the non-parametric methods, it is a parametric probability density function represented as a weighted sum of Gaussian component densities.There is a great similarity between Gaussian mixture model and Vector quantization model in terms of overlapping clusters. The symbol named λ represents collectively these parameters; it is given in formula 2. Each speaker is represented by a GMM and is referred to by his/her model λ.

$$\lambda = \{P_i, \bar{\mu}_i, \Sigma_i\} \quad i = 1, \dots, M \tag{2}$$

where $\bar{\boldsymbol{\mu}}_{\boldsymbol{i}}$ is the mean vector and $\boldsymbol{\Sigma}_{\boldsymbol{i}}$ the covariance matrix of the normally distributed random variable $\boldsymbol{\lambda}$.

In this method, the distribution of the feature vector x is modeled clearly using a mixture of M Gaussians. GMM parameters are estimated from training data using the iterative expectation-maximization (EM) algorithm. These parameters of GMM are computed in training phase to create a speaker model. In testing phase, the speaker model having highest a posteriori probability for the features of an unknown voice is selected as identity of that unknown speaker [14, 17, 43, 44]. A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation:

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i \, g(x \mid \mu_i, \Sigma_i) \tag{3}$$

The following diagram as shown in Figure 4 illustrates the GMM modelling process of speaker data. It shows the illustrative steps of Gaussian mixture modelling of speaker's database.

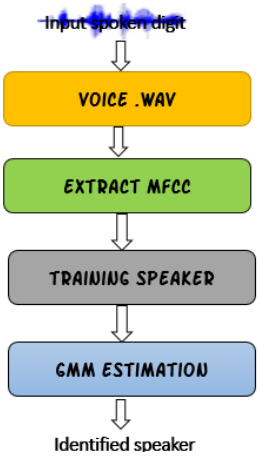

Figure 4. GMM Block diagram of Speaker identification

2.6. Signal to noise ratio

The Signal to Noise Ratio (SNR) is a method to measure the signal strength relative to background noise levels. The SNR is expressed by decibels (dB) using this formula:

$$SNR(dB) = 10 \times log_{10}\left(\frac{P_{speech}}{P_{noise}}\right) \tag{4}$$

where, $P_{speech} = P_{x(t)}$, $P_{noise} = P_{n(t)}$) denote the power of speech signal and noise, respectively. The clean speech signal *x(t)* is degraded by additive signal noise *n(t)* by:

$$n(t) = awgn\ (x,\ snr) \tag{5}$$

So, the observed noisy speech *y(t)* can be expressed as:

$$y(t) = x(t) + n(t) \tag{6}$$

An example of the corruption of the clean signal with an additive white Gaussian noise (AWGN) is given in Figure 5.

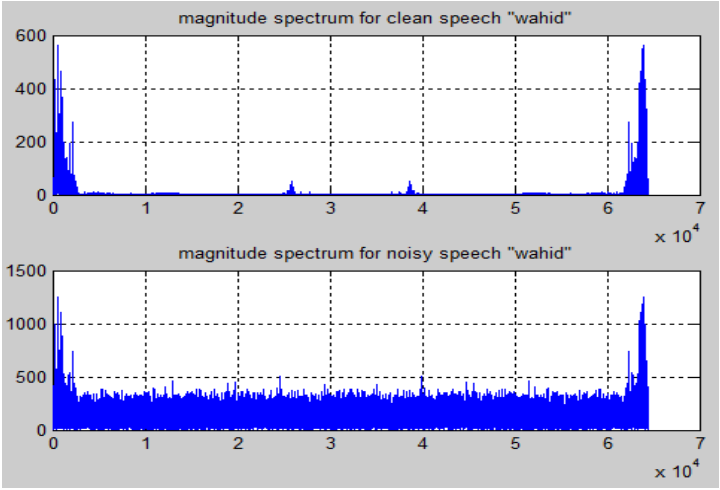

Figure 5. The magnitude spectrum form of the original clean and the noisy signal of the word "/wahid/" recorded in SNR=5 dB

3. EXPERIMENTAL RESULTS

3.1. Corpus preparation

In this paper, we have considered two categories of a database, containing the noisy data and clean data;

– Clean database

We have registered a database ARBDIGITS of 100 Arab Moroccan speakers including males and females have been built from 'siffer' (zero) to 'Tisaa' (nine) which are used for training purpose. The speaker speaks the word several times isolated. These voices have been recorded at sampling frequencies 8000 Hz. We use the noise removal tool available in "audacity" software to delete background noise from the original recording and then we get the clean data.

– Noisy database

For the data that we use it in the testing phase, the noise is added by the MATLAB function adding white Gaussian noise (AWGN) to the clean database ARBDIGITS at various levels of signal to noise ratio (SNR) varying from 5 dB to 20 dB. These testing samples consist of 10 male voices and 5 female voices both aged between 15 and 40 years. They recorded the first ten Arabic digits repeated three times. All the previous analysis was made only in one noisy condition (SNRs level from 5 dB into 20 dB and AWGN noise).

Table 3 reports more technical details about ARADIGITS database used in experimental evaluation.

Table 3. Information and condition of the ARBDIGITS corpus used

Process	Description
Participant	100 Speakers (70 Male 30 females)
Environment	Reverberant and two channels-stereo mode.
Words	10 Arabic spoken Digits
Training Set	85 Speakers
Testing Set	15 Speakers
Number of clean words selected	$10\times3\times100=3000$
Total Number of noisy words	$3000\times1\times4=12000$
noise type used	AWGN
SNR level used	Clean,5 dB, 10 dB, 15 dB, 20 dB
Total Size of Database	1 GB
Sampling Frequency, fs	8000Hz
Software used for mixing	MATLAB R2016b v Trial

3.2. Results

Firstly, the performance of the system is evaluated by recognition rate. It is calculated by:

$$Rec\ Rate = \frac{Successfully\ detected\ word}{Total\ no.of\ words\ in\ test\ dataset} \times 100 \quad (7)$$

The average recognition rates obtained from the ten Arab digits in clean environment and in noisy environment, for different SNR values, are represented in Table 4 and efficiency chart is shown in Figure 6 respectively. From the results shown in Table 4, we can conclude that the effect of the noise is not important if the SNR is superior to 20dB, in this case we obtain approximately the same 'average' value of recognition obtained in cleaned data. In testing phase for speaker identification, the spoken samples are recorded by 15 speakers; (10 male speakers and 5 female speakers) chosen from our database ARBDIGITS of 100 Arab Moroccan speakers (the speech wave is with 8 KHz sampling frequency using AUDIORECORD function of MATLAB 2016 environment in windows platform in 64 bit). The sample collection process is accomplished by using the microphone to record the speech of male/female.

The first testing phase, in clean condition, after this, we have tested with adding AWGN noise at different SNRs levels values of 5, 10, 15, 20 dB. Note that the speech segment was degraded when SNR<5dB. The percentage recognition of a speaker is given in the Table 5 and the efficiency chart is shown in Figure 7 respectively. The average of 15 speaker's recognition rates obtained with training by GMM in clean environment and in noisy environment, for different SNR values are represented in Table 6 and efficiency chart is shown in Figure 8 respectively. Table 7 shows the overall recognition rates (%) for the speaker identification system using the combination of VQ and GMM algorithms and bar chart plot is shown in Figure 9.

Table 4. Average recognition rate (%) using mfcc+vq

Noise Level Digits	Clean	5dB	10dB	15dB	20dB
0	82.37	45.17	67.27	79.66	80.33
1	80.11	42.34	65.17	75.11	79.28
2	81.54	44.06	62.66	78.09	80.05
3	79.08	40.47	60.87	75.33	78.00
4	85.52	41.52	68.12	77.18	84.11
5	80.17	38.78	58.34	74.23	80.17
6	84.42	42.10	66.52	76.12	84.42
7	83.67	39.27	62.71	75.25	82.05
8	77.02	38.96	60.33	71.78	77.02
9	86.63	47.03	61.51	79.00	85.28
Average	82,05	41,97	63,35	76,17	81,07

Table 5. Speaker Identification rate (%) for testing speech in clean and in AWGN noise for different SNR

Methods	#speakers	clean	5dB	10dB	15dB	20dB
MFCC+VQ	10 males	89.43	62.11	77.27	79.66	85.12
	5 females	90.66	63.72	75.17	77.26	86.06

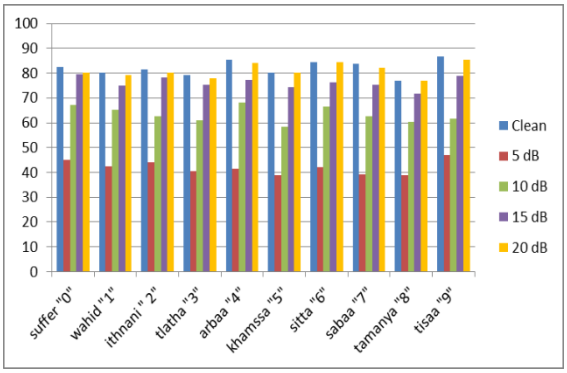

Figure 6. Arab digits recognition success rates (%) in clean and in presence of awgn noise

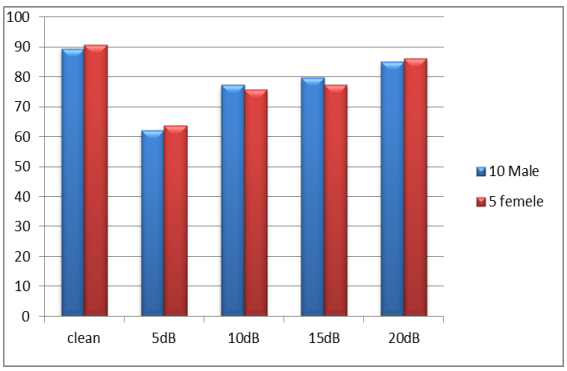

Figure 7. Speaker Identification result (%) in clean and for different SNR values

Table 6. Speaker Identification rate (%) for testing speech in clean and in AWGN noise for different snr using GMM

Methods	#speakers	clean	5dB	10dB	15dB	20dB
MFCC+GMM	10 males	96.12	70.26	88.33	92.66	95
	5 females	94.66	70	87.67	92.33	96.33

Table 7. Speaker Identification rate (%) for testing speech in clean and with AWGN for different snr using GMM +VQ

	Clean	5dB	10dB	15dB	20dB
10 Male	98.33	88.66	90.26	95.12	98
5 female	97.12	86.33	90.66	94	97

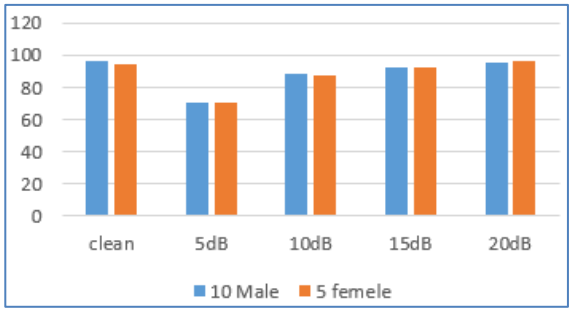

Figure 8. Speaker Identification result (%) in clean and all SNR levels with AWGN noise using GMM modeling

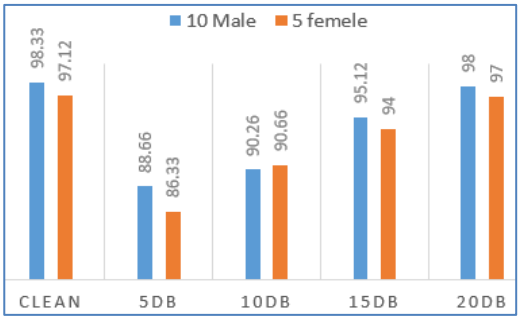

Figure 9. Speaker Identification rate in clean and noise using GMM +VQ

The results are implemented under MATLAB R2016b. For this, we have built a GUI interface as illustrated in Figure 10 to simplify the testing process where the speaker can be tested directly by a new voice recording or from test base. During recording, the user adds AWGN noise to her\his voice if he\she desired and he\she selects the SNR level. This GUI enables the recording or plotting of a sound as well as the recording of a new test data and identification of the speaker ID.

Figure 10. GUI Main MATLAB system

4. DISCUSSIONS

In this work, for the experiments tested with the clean data the maximum performance received for MFCC+GMM+VQ is 97.92%, for MFCC+GMM is 95.39% % and for MFCC+VQ is 90.04 %. Also, for the tests with noisy data, the maximum accuracy received for MFCC+GMM+VQ is 92.50 %, for MFCC+GMM is 86.57% and for MFCC+VQ is 65.64 %. It is clearly observed that better performance has been seen when using the three techniques together MFCC+GMM+VQ. From the results, it is clearly found that the Arabic digits speech recognition system and the speaker identification system performed well in both clean and noisy environment using MFCC as feature extraction and vector quantization. For our proposed model of he combination among three methods MFCC+GMM+VQ, we observed that the accuracy of the results obtained

is high with either clean or noisy data compared to the results obtained by MFCC+GMM or MFCC+VQ methods.We can also conclude that the effect of the noise is not important if the SNR is superior to 20 dB, in this case we obtain approximately same 'average' value of recognition obtained in cleaned data.

Note that the speaker identification system is more performant using the GMM method compared to VQ method, but when we have used the combination of both methods (GMM+VQ) we obtain better speaker identification rate than individual models. Also, we remark an increase of the identification rate by 14% is obtained when SNR=15 dB and GMM as modelling method, but this rate is reduced to 7% when SNR=5 dB. We can explain the difference between those two percentages in the two experiments by the background noise affect on the feature vectors and on the acoustic data, especially if the noise intensity is very strong then it can hide data, consequently the verification is failed. Thus, the verification is performed only on smaller amount of valuable data. As for the Arabic spoken digits system using VQ method, the parameters obtained after training the system for digit 6 and 7 are too close as shown in Figure 7. Therefore, if the digit is not spoken clearly during recognition, the system falters. The digit 8 gives the lowest accuracy, the reason being the speech sample for 8 has the highest amount of "unvoiced" speech signal. Therefore, it is treated as unvoiced speech data.

5. CONCLUSION

In this paper, we have presented an automatic system able to recognize the speaker as well as speech using MFCC, VQ and GMM technique for Arab digits words. The result shows that average accuracy for the system is 90.04% in clean environment and 75.86% in noisy environment for speaker identification using VQ and for Arabic digits recognition system is 82.05% in clean environment and average of 65.64% in noisy environments respectively. The average for speaker identification using GMM is 95.39% in clean environment and 86.57% in noisy environment. For the average of the combination (GMM+VQ) is 97.72% in clean environment and 92,50% adding AWGN noise, so this combination gives better identification rate than individual models. We can improve the obtained results if we use other methods such as ANN, HMM or DNN for classification. In the future works, we will compare this method to other methods in order to find a method that can improve the robustness in other types of noises, testing using other techniques and by increasing vocabulary size.

ACKNOWLEDGEMENTS

The authors thank the reviewers for their suggestions and corrections to the original manuscript. The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] R. G. Mayur, D. Kinnal, and S. B. Ninad, "Classification Techniques for Speech Recognition: A Review," *International Journal of Emerging Technology and Advanced Engineering*, pp. 58-63, 2015.
[2] D. Nikita and Badadapure P. R., "Hindi Speech Recognition System using MFCC and HTK Toolkit," *IJESRT Journal*, vol. 5, no. 12, pp. 90-5, 2016.
[3] N. Singh, A. Agrawal, and R. A. Khan, "A Critical Review on Automatic Speaker Recognition," *Science Journal of Circuits, Systems and Signal Processing*, vol. 4, no. 2, pp. 14-19, 2015.
[4] G. Biagetti, P. Crippa, L. Falaschetti, S. Orcioni, and C. Turchetti, "Speaker Identification in Noisy Conditions Using Short Sequences of Speech Frames," *International Conference on Intelligent Decision Technologies*, pp. 43-52, 2017.
[5] M. Dua, R. K. Aggarwal, and M. Biswas, "Performance evaluation of Hindi speech recognition system using optimized filterbanks," *Eng. Sci. Technol. Int. J.*, vol. 21, no. 3, pp. 389-398, Jun. 2018,
[6] G. Sárosi, M. Mozsáry, P. Mihajlik, and T. Fegyó, "Comparison of feature extraction methods for speech recognition in noise-free and in traffic noise environment," *2011 6th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–8, 2011.
[7] Bhadragiri Jagan Mohan and Ramesh Babu N., "Speech recognition using MFCC and DTW," *2014 International Conference on Advances in Electrical Engineering (ICAEE)*, pp. 1-4, 2014.
[8] T. K. Das and Khalid Nahar, "A Voice Identification System using Hidden Markov Model," *Indian Journal of Science and Technology*, vol. 9, no. 4, pp. 1-6, 2016.
[9] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings IEEE,* pp. 257–286, 1989.
[10] R. Togneri and D. Pullella, "An Overview of Speaker Identification: Accuracy and Robustness Issues," *IEEE Circuits Syst. Mag.*, vol. 11, no. 2, pp. 23–61, 2011.
[11] D. Nagajyothi and P. Siddaiah, "Voice Recognition Based on Vector Quantization Using LBG," *Computer Communication, Networking and Internet Security: Proceedings of IC3T*, pp. 503-511, 2016.
[12] A. Gupta and H. Gupta, "Applications of MFCC and Vector Quantization in speaker recognition," in *2013 International Conference on Intelligent Systems and Signal Processing (ISSP)*, pp. 170-173, 2013.

[13] A. Maurya, D. Kumar, and R. K. Agarwal, "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach," *Procedia Comput. Sci.*, vol. 125, pp. 880-887, 2018.
[14] K. V. Veena and D. Mathew, "Speaker identification and verification of noisy speech using multitaper MFCC and Gaussian Mixture models," *2015 International Conference on Power, Instrumentation, Control and Computing (PICC)*, pp. 1-4, 2015.
[15] Musab T. S. Al-Kaltakchi, Wai L. Woo, Satnam Dlay, and Jonathon A. Chambers, "Evaluation of a speaker identification system with and without fusion using three databases in the presence of noise and handset effects | SpringerLink," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, pp. 1-17, 2017.
[16] Manpreet Kaur and Puneet Mittal, "Speaker Recognition Based on Feature Extraction in Clean and Noisy Environment,"*Journal of Computer Science and Technology,* vol. 25, no. 4, pp. 783-792, 2018.
[17] S. B. Dhonde and S. M. Jagade, "Comparison of Vector Quantization and Gaussian Mixture Model using Effective MFCC Features for Text-independent Speaker Identification," *Int. J. Comput. Appl. Found. Comput. Sci. FCS NY USA*, vol. 134, no. 15, pp. 11-13, 2016.
[18] U. G. Patil, S. D. Shirbahadurkar, and A. N. Paithane, "Automatic Speech Recognition of isolated words in Hindi language using MFCC," *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, pp. 433-438, 2016.
[19] J. Ling, S. Sun, J. Zhu, and X. Liu, "Speaker Recognition with VAD," *2009 Second Pacific-Asia Conference on Web Mining and Web-based Application*, pp. 313-315, 2009.
[20] Yaxin Zhang and M. Alder, "An improved HMM/VQ training procedure for speaker-independent isolated word recognition," *Proceedings of ICSIPNN '94. International Conference on Speech, Image Processing and Neural Networks*, pp. 722-725, 1994.
[21] S. Karpagavalli, R. Deepika, P. Kokila, K. Usha Rani, and E. Chandra, "Isolated Tamil Digit Speech Recognition Using Template-Based and HMM-Based Approaches," *International Conference on Computing and Communication Systems,* 2011.
[22] R. Djemili, H. Bourouba, and M. C. A. Korba, "A speech signal-based gender identification system using four classifiers," *2012 International Conference on Multimedia Computing and Systems*, pp. 184–187, 2012.
[23] Bidhan Barai, Debayan Das, Nibaran Das, Subhadip Basu, and Mita Nasipuri, "An ASR system using MFCC and VQ/GMM with emphasis on environmental dependency," *2017 IEEE Calcutta Conference (CALCON)*, 2017.
[24] Chen Wang, Zhenjiang Miao, and Xiao Meng, "Differential MFCC and Vector Quantization Used for Real-Time Speaker Recognition System," 2008 Congress on Image and Signal Processing, 2008.
[25] N. Hammami and M. Bedda, "Improved tree model for arabic speech recognition," *2010 3rd International Conference on Computer Science and Information Technology*, 2010.
[26] A. Mahmood, M. Alsulaiman, and G. Muhammad, "Automatic Speaker Recognition Using Multi-Directional Local Features (MDLF)," *Arab. J. Sci. Eng.*, vol. 39, no. 5, pp. 3799-3811, 2014.
[27] Amer M. Elkourd, "Arabic Isolated Word Speaker Dependent Recognition System', Islamic University, Gaza, Palestine," vol. 14, no. 1, pp. 15, 2014.
[28] Mansour Alsulaiman, Awais Mahmood, and Ghulam Muhammad, "Speaker recognition based on Arabic phonemes," *Speech Commun.*, vol. 86, no. C, pp. 42-51, 2017.
[29] M. O. M. Khelifa, Y. M. Elhadj, Y. Abdellah, and M. Belkasmi, "Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system," *Int. J. Speech Technol.*, vol. 20, no. 4, pp. 93-949, 2017.
[30] L. Cong and S. Asghar, "Robust Speech Recognition Using Fuzzy Matrix/Vector Quantisation, Neural Networks and Hidden Markov Models," *International Conference on Advances in Pattern Recognition*, London, pp. 255–264, 1999.
[31] F. Merazka, "VQ Codebook Design Using Genetic Algorithms for Speech Line Spectral Frequencies," *International Symposium on Intelligence Computation and Applications*, pp. 557-566, 2012.
[32] A. Samal, D. Parida, M. R. Satapathy, and M. N. Mohanty, "On the Use of MFCC Feature Vector Clustering for Efficient Text Dependent Speaker Recognition," *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA*, pp. 305–312, 2013.
[33] S. Ananthi and P. Dhanalakshmi, "SVM and HMM Modeling Techniques for Speech Recognition Using LPCC and MFCC Features," *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*, pp. 519-526, 2014.
[34] A. Ouisaadane, S. Safi, and M. Frikel, "English Spoken Digits Database under noise conditions for research: SDDN," *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, pp. 1–5, 2019.
[35] A. Touazi and M. Debyeche, "An experimental framework for Arabic digits speech recognition in noisy environments," *Int. J. Speech Technol.*, vol. 20, no. 2, pp. 205-224, 2017.
[36] A. Allosh, N. Zlitni, and A. Ganoun, "Speech Recognition of Arabic Spoken Digits," *Conference Papers in Medicine,* 2013.
[37] H. Satori, M. Harti, and N. Chenfour, "Arabic Speech Recognition System Based on CMUSphinx', *2007 International Symposium on Computational Intelligence and Intelligent Informatics*, pp. 31–35, 2017.
[38] A. I. Amrous, M. Debyeche, and A. Amrouche, "Robust Arabic speech recognition in noisy environments using prosodic features and formant," *Int. J. Speech Technol.*, vol. 14, no. 351, 2011.
[39] S. Satyanand and R. E.g, "Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC," *Int. J. Comput. Appl.*, vol. 17, no. 1, pp. 1-7, 2011.
[40] K. A. Patel, "Speech Recognition and Verification Using MFCC & VQ," *Computer Sience,* 2013.
[41] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*, pp. 248–251, 2012.

[42] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.
[43] Reynolds D., "Gaussian Mixture Models | SpringerLink," *Encyclopedia of Biometrics*, 2009.
[44] D. Desai and M. Joshi, "Speaker Recognition Using MFCC and Hybrid Model of VQ and GMM," in *Recent Advances in Intelligent Informatics*, vol. 235, pp. 53–63, 2014.

BIOGRAPHIES OF AUTHORS

Abdelkbir ouisaadane received a B.Sc. degree in mathematics and Computer Science and an M.Sc. degree in applied mathematics from the faculty of Science and Technics Beni Mellal, Morocco,in 2010 and 2014, respectively. He is currently working towards the Ph.D. degree in Computer Science and Signal Processing from Sultan Moulay Slimane University, Morocco. His research interests include speech and speaker recognition, noise robust signal processing, Arabic speech processing, automatic speech recognition in noisy and reverberant environments, spoken language systems ,machine learning, noise robustness, HMM and neural models for speech applications. E-mail: Abdelkbir.wiss@gmail.com

Said Safi received his B.Sc. degree in Physics (Electronics) from Cadi Ayyad University, Marrakech, Morocco in 1995, M.Sc. degree from Chouaib Doukkali University and Cadi Ayyad University, in 1997 and 2002, respectively. He served a Professor of information theory and telecommunication systems at the National School for applied Sciences, Tangier, Morocco, from 2003 to 2005.Since 2006, he has been a Professor of applied mathematics and programming at Polydisciplinary Faculty, Sultan Moulay Slimane University, Beni Mellal, Morocco. In 2008 he received a Ph.D. degree in Telecommunication and Informatics from the Cadi Ayyad University. In 2015 he received the degree of Professor in Sciences at Sultan Moulay Slimane University. His general interests span the areas of communications and signal processing, estimation, time series analysis, and system identication subjects on which he has published 35 journal papers and more than 70 conference papers. Current research topics focus on transmitter and receiver diversity techniques for single- and multi-user fading communication channels, and wide-band wireless communication systems.
E-mail: sa.said@gmail.com

Miloud Frikel received his Ph.D. degree from the Center of Mathematics and Scientic Computation CNRS URA 2053, France, in array processing. Currently, he is with the GREYC laboratory (CNRSURA 6072) and the ENSI-CAEN as an Assistant Professor. From 1998 to 2003, he was with the Signal Processing Lab,Institute for Systems and Robotics, Institute Superior Tecnico, Lisbon, as a researcher in the eld of wireless location and statistical array processing, after being a research engineer in a software company in Munich, Germany. He worked at the Institute for Circuit and Signal Processing of the Technical University of Munich. His research interests span several areas, including statistical signal and array processing, cellular geolocation (wireless location), space time coding, direction nding and source localization, blind channel identication for wireless communication systems, and MC-CDMA systems. E-mail: mfrikel@greyc.ensicaen.fr