Tunisian Dialect Recognition Based on Hybrid Techniques

Mohamed Hassine, Lotfi Boussaid, and Hassani Massaoud
Laboratoire de Recherche ATSI, Ecole Nationale d'Ingénieurs de Monastir, Tunisia

Abstract: *In this research paper, an Arabic Automatic Speech Recognition System is implemented in order to recognize ten Arabic digits (from zero to nine) spoken in Tunisian dialect (Darija). This system is divided in two main modules: The feature extraction module by combining a few conventional feature extraction techniques, and the recognition module by using Feed-Forward Back Propagation Neural Networks (FFBPNN). For this purpose, four oral proper corpora are prepared by five speakers each. Each speaker pronounced the ten digits five times. The chosen speakers are different in gender, age and physiological conditions. We focus our experiments on a speaker dependent system and we also examined the case of speaker independent system. The obtained recognition performances are almost ideal and reached up to 98.5% when we use for the feature extraction phase the Perceptual Linear Prediction technique (PLP) followed firstly by its first-order temporal derivative (ΔPLP) and secondly by Vector Quantization of Linde-Buzo-Gray (VQLBG).*

Keywords: *Vector Quantization (VQLBG), Mel Frequency Cepstral Coefficients (MFCCs), Feed-Forward Back Propagation Neural Networks (FFBPNN), Speaker Dependent System.*

Received April 24, 2015; accept February 3, 2017

1. Introduction

Arabic is the sixth most widely spoken language in the world. It is the language of over 24 countries and spoken by more than 300 million of persons. In Arabic language, we distinguish three varieties: The Classical Arabic (CA) that is the language of the Koran, the Modern Standard Arabic (MSA), which is used in media and studied at school, and the Dialect [10].

There are two major groups in Arabic dialect: the western Arabic dialect (Maghreb or North Africa) and the eastern one (Levantine Arabic, Gulf Arabic and Egyptian Arabic). Here, we focus on the Tunisian Dialect (TD) which belongs to western Arabic. In Tunisia, the dialect is called 'Darija' or 'Tounsi', which means Tunisian, in order to be distinguished from the literal Arabic [5]. 'Darija' is the daily spoken language; it is different from the CA and the MSA. It has two important forms that are the urban dialect and the rural dialect. There are also some regional variations in these dialects: The variety of Tunis region, the Sahelian variety, the Sfaxian, the southern, etc.

In TD, we distinguish the absence of the contingencies endings, modification of the paradigm of the conjugation, different order of words in the sentence, the use of terms borrowed from western languages such as Turkish, Spanish and French that are the languages of the old colonial powers [3]. The lack of researches on automatic Arabic speech recognition is noticeable compared to other languages such as English, French, etc. This may be motivated by the reliability of several dialects in Arabic language and by the crucial complexity of such language on different levels: Phonetic, linguistic, semantic, contextual, morpho-syntactic, etc.

In this work, we are interested in automatic recognition of the ten Arabic digits (from zero to nine) pronounced in TD 'Darija'. Arabic digits are interesting and can be considered as representative elements of language because more than half of the phonemes of the Arabic language are included in the 10 digits [16]. The recognition of digits can be the first steps for other Arabic isolated word recognition and continuous speech recognition. The digits recognition has many applications such as facilitating communication for people with functional disability: voice command will be used instead of mechanical command as: dialing a telephone number by voice, airline reservation, banc systems, robotic simulation, oral messaging, etc.

To implement our system, a multitude of hybrid feature extraction techniques are used: The Perceptual Linear Prediction (PLP) technique followed firstly by its first-order temporal derivative (PLP+ΔPLP) and secondly by the Vector Quantization (VQLBG), the Mel Frequency Cepstral Coefficients technique followed firstly by the first-order temporal derivative (MFCC+ΔMFCC) and secondly by VQLBG. The latters were the main interesting feature extraction techniques used here that were joined to FFBPNN for the recognition phase.

Our present system succeeded in reaching good performances, which exceeded 98.5% in many cases.

This paper is organized as follows: in section 2, we shed some lights on basic related works. In section 3, we describe some theoretical backgrounds. The suggested methodology is detailed in section 4, while section 5 describes experimental results. The main results are discussed in section 6 and finally conclusion and future works are given in section 7.

2. Related Works

TD is classified among the under-resourced languages. It has neither a standard orthographic or written text nor dictionaries. Consequently, researches on TD recognition are limited and are interested in restricted fields. In Tunisian Railway Transport Network domain, Masmoudi *et al.* [19] have built an automatic TD recognition system by creating a corpus named “TARIC” and a pronunciation dictionary based on a list of graphemes, phonemes, lexicon of exceptions and phonetic rules. The prepared dictionary was evaluated on two types of corpora. The word error rate of word grapheme-to-phoneme mapping was around 9%. In the same field of Railway Transport Network, TD was investigated by [14]. Their work consists of building a recognition system for semantic annotation and semantic interpretation of Tunisian utterances based on ontology. The proposed method is tested on a TD corpus. The obtained accuracy is about 0.96.

Boujelbane *et al.* [5, 6] have developed an approach which consists in studying the morphological, syntactic and lexical differences between MSA and TD by exploiting the Penn Arabic tree bank. The knowledge and relationships between TD and MSA were operated to describe a method for building a bilingual dictionary and creating TD corpora.

Zribi *et al.* [25] have proposed a method to adapt an MSA morphological analyzer for the TD by exploiting the points of similarities (between MSA and TD). The test performance achieved encouraging results: an F-measure of 88%.

Because of the importance of the digits recognition in term of social applications, El-Mashed *et al.* [11] have been interested on connected Arabic digits (numbers) where independent speaker Arabic speech recognition is used in order to recognize Colloquial Egyptian dialect.

The proposed approach is divided into four stages:

- Segmentation of each pronounced number in ten digits.
- Feature extraction which is consisted of the Mel Frequency Cepstral Coefficients of these digits;
- Application of K-means clustering algorithm for the latter features in order to extract the relevant information;
- And finally the use of the Support Vector Machine (SVM).

This approach yields to 94% accuracy.

Previously, Ganoun and Almerhag [13] have developed a system for recognizing spoken Arabic digits from zero to nine based on three feature extraction techniques: Yule-Walker spectrum feature, Walsh spectrum feature and Mel Frequency Cepstral Coefficients. It was found that the MFCC provides the best recognition rate, while the worst rate was that of Yule-Walker.

Recently, [7] have been interested in automatic recognition of Arabic digits from zero to nine uttered by 24 speakers in three Arabic dialects: Egyptian, Jordanian and Palestinian. The feature extraction has been realized by combining wavelet transform with the linear prediction coding and the classification by Probabilistic Neural Network (PNN). The average recognition rate reached 93%. The recognition performance in noisy environment has been also investigated and the obtained results were very promising.

In [4], mono-speaker speech recognition of 11 Arabic words is realized. The authors used the MFCC followed by Bionic Wavelet Transform (BWT) for feature extraction. In the classification phase, Feed-Forward Back Propagation Neural Network (FFBPNN) is used. With this system, the recognition rate reached 89.09% with MFCC followed by BWT and 99.39% with the second derivative of MFCC followed by BWT ($\Delta\Delta$MFCC+BWT).

In [1], a system of automatic Arabic word recognition is proposed where the effectiveness of discrete wavelet transform is experienced. It was proved that neural network embedded with wavelet yields a good recognition result with 77% accuracy.

Salam *et al.* [22] have proposed a Malay isolated speech recognition system using neural network. In this work, various experiments were conducted to choose suitable number of nodes in hidden layer and learning parameters for the feed forward multilayer perceptron system. Best recognition rate achieved was 95% using network topology of input nodes, hidden nodes and output nodes of size 320:45:4 respectively [22].

3. Theoretical Background

3.1. Features Extraction

To extract the relevant information, to minimize noise and to remove the redundancy from the speech, several features extraction methods are needed and have been used separately and jointly as: PLP, ΔPLP, MFCC, ΔMFCC, Relative Spectral Perceptual Linear Prediction (RASTA-PLP), Continuous Wavelet Transform (CWT), Digital Wavelet Transform (DWT). In order to reduce dimensionality of the obtained features the latters were always followed separately by other algorithms such as: K-Means, Fuzzy Clustering Means (FCM), Principal Component Analysis (PCA) and VQLBG. In this section, we will detail those that have realized credible results.

3.1.1. The Mel Frequency Cepstral Coefficients

The MFCCs are dominant in speech recognition areas; this feature extraction technique uses a non-linear frequency scale, which is the Mel scale, in order to simulate the frequency response of the human auditory system. MFCCs are based on known variation of the human ear's critical bandwidth with frequency [20, 21]. It is a psychoacoustic measure of pitches judged by human that is linear in bottom of 1000Hz [23] and logarithmic above. The MFCCs provide a compact representation of the given speech signal. The mathematical relationship between Mel frequency scale and linear frequency scale is defined by:

$$f_{Mel} = 2595 \times \log(1 + f_{HZ} / 700) \qquad (1)$$

Where f_{HZ} is the frequency in Hz.

To compute MFCCs we have used the following steps:

a. Pre-emphasis: each signal corresponding to each digit is pre-emphasized to increase the contribution of the high frequencies in the speech signal:

If $s(n)$ is the original speech signal and $s_p(n)$ is the pre-emphasized signal then:

$$s_p(n) = s(n) - 0.97s(n-1) \qquad (2)$$

This implies to filter the speech signal in a Finite Impulse Filter Response (FIR) whose transfer function in Z domain is [12]:

$$h_p(z) = 1 - 0.97z^{-1} \qquad (3)$$

b. Windowing: in this stage, the pre-emphasized signal is divided into frames of 25ms each, which corresponds to 25 10^{-3} 44100=1102 samples, and multiplied by an overlapped sliding Hamming window with an overlapping step of 10ms (441 samples) to avoid leakage and spectral distortion at the beginning and at the end of each frame. The Hamming window is given by [11].

$$h(n) = \begin{cases} 0.54 - 0.46\cos(\dfrac{2\pi n}{N-1}) & if \quad 0 \le n \le N-1 \\ 0 \quad otherwise & \end{cases} \qquad (4)$$

Where N is the number of samples in the window.

c. Discrete Fourier Transform (DFT): DFT is used to convert each frame of N samples from the time domain to the frequency domain, which yields to the signal spectrum:

$$S(k) = \sum_{n=0}^{N-1} s(n)e^{-j2\pi kn/N} \qquad k = 0,1,2,\ldots,N-1 \qquad (5)$$

d. Mel filters bank: since the frequencies range obtained in the previous step is wide, to avoid huge calculations, a filter bank in the Mel scale is built to pass the speech signal through it. The Mel filters bank are series of overlapped triangular filters, which are built in such a way that the low boundary of a filter is situated at the center of the previous filter and the upper boundary is at the next filter.

Assume that $H_m(k)$ is the frequency magnitude response of the m^{th} filter of Mel filter bank, where k is the discrete frequency index in the digital domain. The filter output of the m^{th} filter, X_m, can be expressed by:

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S(k)|^2 |H_m(k)| \qquad 1 \le m \le M \qquad (6)$$

M is the total number of filters.

e. Discrete Cosine Transform (DCT): in this step, the DCT is done, which yields to the Mel Cepstral Coefficients defined by: $c(m)=DCT(log(X_m))$

f. The first-order temporal derivative coefficients of MFCCs(ΔMFCCs):

ΔMFCCs are also known as differential coefficients. These correspond to the trajectories of the basic MFCCs coefficients and represent their variability in time. ΔMFCCs are computed by the following regression equation [24]:

$$d_i = \frac{\sum_{n=1}^{N} n(c_{n+i} - c_{n-i})}{2\sum_{n+1}^{N} n^2} \qquad (7)$$

Where d_i is the delta coefficient at frame i computed in terms of the corresponding basic Cepstral Coefficients C_{n+i} to C_{n-i}. A typical value for N is 2.

3.1.2. Perceptual Linear Prediction Coefficients

The PLP is another feature extraction technique, which emulates the human auditory system and uses a Bark scale that is different from the Mel scale used in MFCCs. There are three main concepts behind PLP [15]. They are critical band frequency selectivity, equal-loudness curve and intensity-loudness power law.

The relationship between the Bark frequency scale and the linear frequency scale is given by:

$$f_{bark} = 6\ln\left(\tfrac{f}{600} + \left(\left(\frac{f}{600}\right)^2 + 1\right)^{0.5}\right) \qquad (8)$$

f is the frequency in Hz.

To compute PLP coefficients, the steps below are followed:

The three first steps are similar to that of the MFCCs, the difference here is the use of a filter bank in Bark scale instead of Mel scale and the remaining steps are:

a. Equal-loudness curve: The role of equal-loudness curve is to approximate the sensitivity of human hearing at various different frequencies.

Assume that $S(k)$ is the signal spectrum at frequency index k, and $\psi_m(k)$ are the filter weights of the m^{th}

Bark filter along the discrete linear frequency scale, so the filter output of the m^{th} filter is:

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} \left|S(k)\right|^2 \left|\psi_m(k)\right| \qquad 2 \le m \le M-1 \tag{9}$$

Let E_m be the equal loudness weight of the $\mathrm{m_{th}}$ filter:

$$E_m = \frac{(\omega^2 + 56.8\times10^6)\omega^4}{(\omega^2+6.3\times10^6)^2(\omega^2+0.38\times10^9)} \qquad \omega = 2\pi f \tag{10}$$

$X_{m(e)}=E_m X_m$; is the m^{th} filter output after applying the equal loudness weight.

b. Intensity-loudness power law: in this step, the non-linear relationship between signal intensity and perceived loudness is described.

Mathematically it is expressed by the following formula [18]:

$$\phi_m = (X_{m(e)})^{0.33} \qquad 1 \le m \le M \tag{11}$$

c. Inverse Discrete Fourier Transform (IDFT) and Cepstral analysis: here, an IDFT is applied to the filter outputs ϕ_m, then Levinson-Durbin Algorithm is applied to the obtained result to compute the Linear Prediction Coefficients (LPC), finally Cepstral Coefficients $\hat{v}[n]$ are computed by applying the following formula [17]:

$$\begin{aligned} &\hat{v}[n] = \ln(G) \qquad for \quad n=0 \\ &\hat{v}[n] = a_n + \sum_{k=1}^{n-1}\left(\frac{k}{n}\right)\hat{v}[k]a_{n-k} \qquad for \quad 1 \le n \le p \end{aligned} \tag{12}$$

Where $\hat{v}[n]$ is the $\mathrm{n^{th}}$ order Cepstral Coefficient, G the LPC filter gain, P is the LPC filter order and a_n is the $\mathrm{n^{th}}$ order linear prediction coefficient.

3.1.3. Vector Quantization (VQ)

The vector quantization is a process of mapping vectors from a vector space to a finite number of regions in that space. Here the LBG algorithm is used and implemented by the following recursive procedure:

1. Design a 1-vector codebook: this is the centroid of the entire set of training vectors (hence, no iteration is required here).
2. Double the size of the codebook by splitting each current codebook Y_n according to the rule:

$$Y^+_n = Yn(1+\varepsilon)$$
$$Y^-_n = Yn(1-\varepsilon)$$

Where n varies from 1 to the current size of the codebook, and ε is a splitting parameter (we choose ε=0.01).

3. Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).
4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.
5. Repeat steps 3 and 4 until the average distance falls below a preset threshold.
6. Repeat steps 2, 3, and 4 until a codebook size of M is designed [2].

3.2. Recognition Phase:

In the recognition phase, the Artificial Neuronal Network (ANN) is used. ANN architecture is a simulation of information processing that occurs in the biological brain. It starts with receiving, learning, adapting, recognizing the pattern and performing a desired function (target) by trial of different weights of the information elements in a computation model. A typical ANN model contains an input layer that receives the input data. The hidden layers with number of nodes that would satisfy the problem requirement would recognize patterns and organize these data through multiple trial processes to predict the output [8]. In our work, the FFBPNN is used. In this ANN type, neurons are connected forward where each layer of the neural network connects to the next layer. Here, the FFBPNN consists of an input layer, one hidden layer and one output layer.

4. Methodology

The methodology of our work is performed as the following steps:

a. Recording and preprocessing: the recording of the ten digits has been occurred in suitable conditions where professional acoustical materials:

A digital mixing console (Studer on air 2000 M2), a dynamic microphone (MD 421) were used to capture the speech signal wave, a professional software (Sound Forge 6.0) for recording, cleaning and organizing the digits in separate files. The speech was recorded in Mono wave files, at a sampling rate of 44100 Hz and coded in 16 bits.

Four proper corpora were prepared by 10 voluntary speakers (5 males and 5 females) aged between 9 and 60 years, each speaker pronounced each digit five times. The first corpus is a mixture. It includes the speech signal of five speakers (3 males and 2 females).

The second is also a mixture. It is built with the speech signal of the remaining five speakers (2 males and 3 females). In the third corpus, we grouped all the males of the two previous corpora and the fourth corpus is composed of the speech signal of all the females of the two previous mixture corpora. This diversity of corpora is used to validate the accuracy and the final system performances.

The training database for each corpus mentioned above is composed of the first four trials of the five speakers and it counts 200 files (20 files for each digit: composed of four files for each speaker). The validation database contains the first trial of the five speakers meanwhile the testing database is composed of the fifth trial of each speaker and counts 50 files (5 files for each digit: composed of one file for each speaker). Therefore, the training database is composed of 80% of the original corpus, the validation database is composed of 20% of the original corpus from that has been included in the training database, and the testing database is composed of 20% of the original corpus from who has not been included in the training database.

b. Applying the features extraction techniques: after pre-emphasizing the speech signal of each digit, the already mentioned feature extraction techniques were applied separately and sometimes jointly to the recorded speech signal. When MFCCs technique is applied, we used a filter bank of 40 filters where the first thirteen are linear and the remained are logarithmic. One matrix of 13 lines and a variable number of columns is obtained for each digit. After applying ΔMFCCs, a matrix with the same dimension is obtained too. This latter was concatenated with the basic Cepstral Coefficients (MFCCs) matrix of the same digit in order to form one matrix, which represents one digit.

When the PLP and ΔPLP techniques were applied, the same steps were followed. Since the amount of data of each digit matrix is large after applying feature extraction technique, the use of LBG algorithm drastically reduces the dimensionality of features. This can increase the robustness of the recognition system and decrease the computational time and the requirement for large memory calculations.

After concatenating MFCCs and ΔMFCCs or PLP and ΔPLP of the original signal of each digit, the LBG algorithm is used to reduce each feature digit matrix to two columns (2 vectors) and the number of lines is kept (13). Then, the two columns are newly concatenated to form one column vector: so each digit is represented by one vector. These steps were repeated for the speech signal of all the digits.

Finally, we obtained a matrix of 200 vectors for the training database, 50 vectors for the validation database and 50 vectors for the test database: These will be used as inputs for FFBPNN in the recognition phase.

c. Applying FFBPNN: During the training, the input layer is fed with a matrix of 13 lines and 200 columns, which represent 80% of the database corresponding to the digits of the corpus. Each column represents the feature of one digit. During the testing step, the FFBPNN is fed with a matrix of 13 lines and 50 columns. These represent 20% of the database. For the hidden layer we choose a number of neurons always equal to 70 and sometimes 90 then the “TanSig” activation function. For the output layer, we choose seven neurons and the “LogSig” activation function. The neural network has been trained in supervised mode. We used a binary code of 7 bits as a Target. The performance function is Mean Square Error (MSE) and the training function is ‘Trainlm’. The remaining parameters are taken by default.

5. Experimental Results

Our work is conducted in five experiments where features extraction and recognition were implemented in Matlab7.1 platform language. We let Matlab program prepared for our recognition system running until one of the known Multi-Layer Perceptron (MLP) stop criterions is reached, and we note each time the corresponding error rates.

- In the first experiment, we use the first mixture corpus where several feature extraction techniques were experimented as described in Table 1. During the entire experiment, we noticed that MFCCs followed firstly by ΔMFCCs and secondly by VQLBG has realized the best test error rate of 1.41%. The PLP followed firstly by ΔPLP and secondly by VQLBG has occupied the second order in term of performances with 1.46% test error rate. The remaining experimented techniques have realized acceptable results but not satisfactory.
- In the second experiment, the second mixture corpus is used. The best results were obtained by PLP followed firstly by ΔPLP, and secondly by VQLBG and the test error rate reached 1.55% as shown in Table 2. Error rate curves of training, validation and test are shown in Figure 1.
- In the third experiment, the male corpus has been tested. The best performance reached was obtained with PLP followed firstly by ΔPLP and secondly by VQLBG as shown in Table 3. Error rate curves of training, validation and test are shown in Figure 2.
- In the fourth experiment, the female corpus was computed with PLP followed firstly by ΔPLP and secondly by VQLBG. Obtained performance is shown in Table 4. Error rate curves of training, validation and test are shown in Figure 3.
- In the fifth experiment, the speaker independent system is examined. We preserved the training database of the first experimented corpus and we used other speakers who have not participated in the entire previous corpora to build a new test corpus.

We also kept the regular percentage concerning the amount of data in training database, in validation database and in test database. The obtained results are shown in Table 5.

Table 1. Obtained error rates computed on the first mixed corpus.

Technique	Train Error in %	Validation Error in %	Test Error in%	# Epochs	# Neurons in hidden Layer
PLP+PCA	418.956e-016	159.286e-015	2.36	25	90
PLP+ Δ PLP+PCA	0.140954	0.149506	1.61	450	90
DWT +PLP+PCA	155.612e-015	127.251e-015	2.96	23	90
Rasta-PLP+PCA	590.812e-015	114.715e-014	3.99	21	90
MFCC+PCA	194.019e-015	190.624e-015	6.53	27	90
MFCC+ Δ MFCC+PCA	0.132904	0.10428	4.40	370	90
DWT +MFCC+PCA	204.327e-016	138.065e-01	8.42	33	90
MFCC+VQLBG	329.818e-016	104.046e-015	5.17	47	90
MFCC+ Δ MFCC+VQLBG	547.279e-005	556.211e-005	1.41	140	90
PLP+VQLBG	0.00951932	0.00904045	4.04	210	90
PLP+ Δ PLP+VQLBG	869.276e-016	141.301e-015	1.46	49	70
Rasta-PLP+VQLBG	408.416e-015	235.33e-015	2.41	21	70
PLP+K-means	414.876e-016	598.511e-016	13.82	24	70
PLP+ Δ PLP +K-Means	0.085122	0.0758796	7.20	180	70
Rasta-PLP+K-Means	0.285891	0.285988	18.22	24	70
MFCC+K-means	0.603431	0.857169	13.51	76	90
MFCC+ΔMFCC+ k-Means	252.418e-015	118.143e-016	10.23	24	70
PLP+FCM	405.432e-015	415.434e-015	11.76	28	100
PLP+ΔPLP+FCM	154.563e-015	135.241e-015	2.64	79	90
Rasta-PLP+FCM	436.456e-015	226.446e-015	8.75	21	90
MFCC+FCM	125.258e-015	857.676e-016	10.99	37	100
MFCC+ΔMFCC +FCM	256.406e-015	210.684 e-015	6.46	23	90

Table 2. Obtained error rates with the second mixed corpus.

Technique	MFCC + ΔMFCC + VQLBG	PLP + ΔPLP +VQLBG
Training Error in %	392.459e-016	182.131e-015
Validation Error in %	142.297e-016	447.39e-016
Test Error in %	2.27	1.55
# Epochs	24	25
# Neurons in Hidden Layer	70	70

Table 3. Obtained error rates with a female corpus.

Technique	MFCC + ΔMFCC + VQLBG	PLP + ΔPLP +VQLBG
Training Error in %	155.064e-015	687.98e-016
Validation Error in %	106.817e-015	848.96e-016
Test Error in %	2.42	1.84
# Epochs	23	59
# Neurons in Hidden Layer	70	90

Table 4. Obtained error rates with a male corpus.

Technique	MFCC + ΔMFCC + VQLBG	PLP + ΔPLP +VQLBG
Training Error in %	731.284e-016	333.88e-016
Validation Error in %	618.523e-016	469.51e-016
Test Error in %	1.71	0.95
# Epochs	22	34
# Neurons in Hidden Layer	70	70

Table 5. Results with speaker independent system.

Technique	MFCC + ΔMFCC + VQLBG	PLP + ΔPLP +VQLBG
Training Error in %	125.036e-015	270.581e-016
Validation Error in %	100.52e-015	275.669e-016
Test Error in %	11.38	10.20
# Epochs	22	93
# Neurons in Hidden Layer	70	70

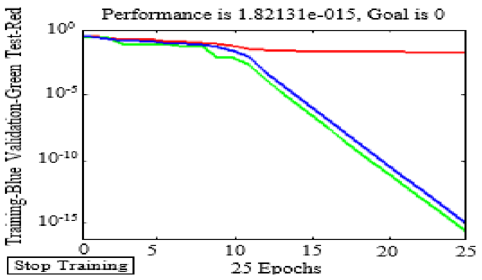

Figure 1. Obtained error rate curves with the second mixed corpus by using PLP+ΔPLP+VQLBG.

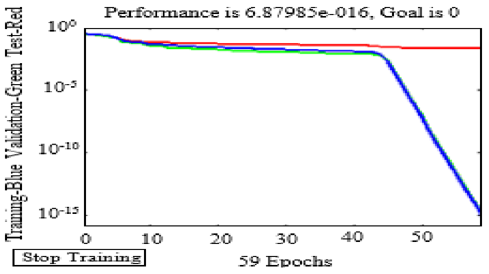

Figure 2. Obtained error rate curve with the male corpus by using PLP+ΔPLP+VQLBG.

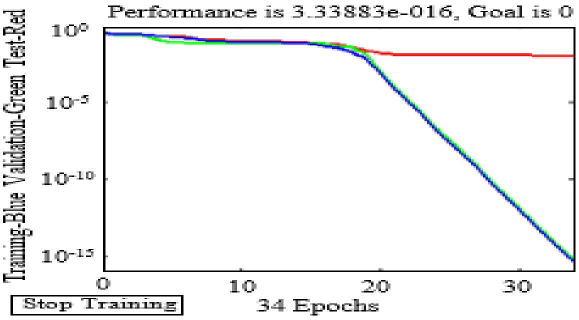

Figure 3. Obtained error rate curves with the female corpus by using PLP+ΔPLP+VQLBG.

6. Results and Discussion

The results presented in Table 1 showed the recognition performances with different feature extraction techniques when using the first corpus. The remaining tables showed the obtained results with the two best feature extraction techniques (MFCCs and PLP) by using different other corpora in each table. The average test performance in these four tables is 98.04% when using MFCCs and 98.54% with PLP. It is clear that PLP is the best feature extraction technique. The results given in reference [7] for Arabic digit recognition based on wavelet transform with the linear prediction coding using PNN showed a recognition performance of 93%. Our proposed system proved that it is better than the system proposed in reference [7]. In reference [9], in which the author is interested in Arabic digit recognition based on the time-frequency analysis of the wavelet transform followed by MFCCs for feature extraction technique and Hidden Markov Model (HMM), the found performance is 98% while it is 98.54% in our system when using PLP. Therefore, the found result with our system outperforms those of found in reference [7, 9]. In our present work, the found test error rates when using a diversity of corpora were close to each other and this may prove and validate the effectiveness of our system. For the speech recognition, it is recommended to use PLP since its performance is found to be 98.54%. The PLP technique adopts three essential properties, which are the integration of critical bands, the equal loudness pre-emphasis, and the intensity-loudness conversion. With these aspects, the PLP becomes nearer to the human hearing than other techniques and consequently it allows obtaining robust and discriminatory parameters.

7. Conclusions and Future Work

In this paper, we proposed a speech recognition system based on hybrid recognition techniques for feature extraction and FFBPNN for classification. In order to validate our work, four corpora were experimented. It was also shown during the entire experiments that using PLP followed firstly by ΔPLP and secondly by VQLBG offer an average test performance of 98.54%. It was too often seen that the LBG algorithm is better than PCA algorithm in term of performances and in computational times in all the experiments. The case of speaker independent system was investigated and it was shown that the obtained results were acceptable. In the future, we plan to expand our database in order to cover all the Arabic dialects, to use more advanced techniques which respect the non-linearity of the speech and to extend our work for continuous Arabic speech recognition and Arabic dialect classification.

References

[1] Al-Irhaim Y. and Saeed E., “Arabic Word Recognition Using Wavelet Neural Network,” *in Proceeding of Third Science Conference in Information Technology*, Al Mosul, pp. 416-425, 2010.

[2] Ameen A., Uma R., and Madhusudana R., “Speaker Recognition System Using Combined Vector Quantization and Discrete Hidden Markov Model,” *International Journal Of Computational Engineering Research*, vol. 2, no. 3, pp. 692-696, 2012.

[3] Baccouche T., *L’emprunt En Arabe Moderne*, Beit El-hikma Et Iblv, 1994.

[4] Ben-Nasr M., Talbi M. and Cherif A., “Arabic Speech Recognition by MFCC and Bionic Wavelet Transform using a Multi-Layer Perceptron for Voice Control,” *CiiT International Journal of Software Engineering and Technology*, vol. 4, no. 3, 2012.

[5] Boujelbane R., Khemekhem M., and Belguith L., “Mapping Rules for Building a Tunisian Dialect Lexicon and Generating Corpora,” *in Proceedings of International Joint Conference on Natural Language Processing*, Nagoya, pp. 419-428, 2013.

[6] Boujelbane R., Ellouze M., Hadrich Belguith L., “De L’arabe Standard Vers L’arabe Dialectal: Projection De Corpus Et Ressources Linguistiques En Vue Du Traitement Automatique De L’oral Dans Les Médias Tunisiens,” *in Proceedings of Tunisia International Joint Conference on Natural Language Processing*, Nagoya, pp. 419-428, 2013.

[7] Daqrouq K., Alfaouri M., Alkhateeb A., Khalaf E. and Morfeq A., “Wavelet LPC with Neural Network for Spoken Arabic Digits Recognition System,” *British Journal of Applied Science and Technology*, vol. 4, no. 8, pp. 1238-1255, 2014.

[8] El-Baroudy I., Elshorbagy A., Carey S., Giustolisi O., Savic D., “Comparison of Three Data-Driven Techniques in Modelling the Evapotranspiration Process,” *Journal of Hydro Informatics*, vol. 12, no. 4, pp. 365-379, 2010.

[9] El-Henawy I., Khedr W., ELkomy O., Abdalla A., “Recognition of Phonetic Arabic Figures Via

Wavelet Based Mel Frequency Cepstrum Using Hmms," *HBRC Journal*, vol. 10, no. 1, pp. 49-54, 2014.

[10] Elmahdy M., Gruhn R., Minker W., and Abdennadher S., "Modern Standard Arabic Based Multilingual Approach for Dialectal Arabic Speech Recognition," *in Proceedings of IEEE 8th International Symposium on Natural Language Processing*, Bangkok, pp. 169-174, 2009.

[11] EL-Mashed S., Sharway M., and Zayed H., "Speaker Independent Arabic Speech Recognition Using Support Vector Machine," *in Proceedings of ICI-11 Conference and Exhibition on Information Technology and Instruction Technology*, Hungary, pp. 401-416, 2011.

[12] Ganchev T., "Speaker Recognition," PHD Theses, University of Patras, 2005.

[13] Ganoun A. and Almerhag I., "Performance Analysis of Spoken Arabic Digits Recognition Techniques," *Journal of Electronic Science and Technology*, vol. 10, no. 2, pp. 153-157, 2012.

[14] Graja M., Jaoua M., and Belguith L., "Building Ontologies to Understand Spoken Tunisian Dialect," *International Journal of Computer Science, Engineering and Applications*, vol. 1, no. 4, pp. 23-32, 2011.

[15] Gunawan W. and Hasegawa-Johnson M., "PLP Coefficients Can be Quantized at 400 BPS," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, 2001.

[16] Hamdi R., "La variation Rythmique Dans Les Dialects Arabes," PhD Thesis, Université Lumière Lyon2 & Université 7 Novembre à Carthage, 2007.

[17] Haykin S., *Neural Networks and Learning Machines*, Prentice Hall, 2009.

[18] Hermansky H., "Perceptual Linear Predictive (PLP) Analysis for Speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, 1990.

[19] Masmoudi A., Khemakhem M., Estève Y., Belguith L., and Habash N., "A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition," *in Proceedings of the 9th International Conference on Language Resources and Evaluation*, Iceland, pp. 306-310, 2014.

[20] Muda L., Begam M., and Elamvazuthi I., "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," *Journal Of Computing*, vol. 2, no. 3, pp. 138-143, 2010.

[21] Price J., "Design an Automatic Speech Recognition System Using Maltab," University of Maryland Eastern Shore Princess Anne, 2005.

[22] Salam M., Mohamad D., and Salleh S., "Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes and Learning Parameters," *The International Arab Journal of Information Technology*, vol. 8, no. 4, pp. 364-371, 2011.

[23] Semet G. and TREFFO G., "Reconnaissance De La Parole Avec Les Coefficients MFCC," *in Proceedings of TIPE*, 2002.

[24] Srinivasan A., "Speech Recognition Using Hidden Markov Model," *Applied Mathematical Sciences*, vol. 5, no. 79, pp. 3943-3948, 2011.

[25] Zribi I., Khemekhem M., and Belguith L, "Morphological Analysis of Tunisian Dialect," *in Proceedings of International Joint Conference on Natural Language Processing*, Nagoya, pp. 992-996, 2013.

Mohamed Hassine has received a Diploma in electrical Engineering in 1997, his Master in 2005 and his PhD degree in Electrical Engineering in 2017 from the National School of Engineering of Monastir, University of Monastir in Tunisia. His current research interests include automatic speech recognition.

Lotfi Boussaid has received a Diploma in Electrical Engineering in 1989 from the University of Monastir in Tunisia, his Master in "Nouvelles Technologies des Systèmes Informatiques Dédiés" in 2003 and his PhD degree in Computer Science in 2006 from the University of Sfax. He was a member of LE2I, the laboratory of Electronic, Computing and Imaging Sciences, Burgundy University, France. His current research interests include Hardware-Software design space exploration and prototyping strategies for real-time systems.

Hassani Messaoud has received his Bachelor's degree in Electrical Engineering in 1983 and his Master of Science in Control Engineering 1985 from the High Normal School of Technical Education (ENSET) in Tunis-Tunisia. His PhD in Control Engineering was prepared at the University of Nice-Sophia Antipolis / France in 1993 and his Habilitation Diploma was defended at the School of Engineers (ENIT) in Tunis -Tunisia. He is presently a Professor at the School of Engineers of Monastir-Tunisia (ENIM). His main interest is robustness in identification and control of non-linear systems with application to diagnosis and equalization of numerical communication channels.