International Journal of Intelligent Information Systems
2015; 4(1): 27-34
Published online February 2, 2015 (http://www.sciencepublishinggroup.com/j/ijiis)
doi: 10.11648/j.ijiis.20150401.14
ISSN: 2328-7675 (Print); ISSN: 2328-7683 (Online)

Hybrid techniques for Arabic letter recognition

Mohamed Hassine[1], Lotfi Boussaid[2], Hassani Massouad[1]

[1]LARATSI Lab, ENIM, University of Monastir, Monastir, Tunisia
[2]EμE Lab, FSM, University of Monastir, Monastir, Tunisia

Email address:
hassinemohamed60@yahoo.com (M. Hassine), lotfi.boussaid@enim.rnu.tn (L. Boussaid), hassani.massaoud@enim.rnu.tn (H. Messaoud)

To cite this article:
Mohamed Hassine, Lotfi Boussaid, Hassani Massouad. Hybrid Techniques for Arabic Letter Recognition. *International Journal of Intelligent Information Systems.* Vol. 4, No. 1, 2015, pp. 27-34. doi: 10.11648/j.ijiis.20150401.14

Abstract: In this paper we investigate the use of the feed-forward back propagation neural networks (FFBPNN) for automatic speech recognition of Arabic letters with their four vowels (Fatha, dhamma, Kasra, Soukoun). This investigation will constitute a basically step for the recognition of continuous Speech. Features were extracted from recorded corpus by using a variety of conventional methods such as Linear Predictive Codes (LPC), Perceptual Linear Prediction (PLP), Relative Spectral Perceptual Linear Prediction (RASTA-PLP), Mel Frequency Cepstral Coefficients (MFCC), Continuous Wavelet Transform (CWT), etc. Here, several hybrid methods have been used too. Since the extracted features have large dimensionalities they were reduced by conserving the most discriminatory information with the Principal Component Analysis (PCA) technique. The recognition performance has been improved particularly when we use the PLP method followed by PCA technique.

Keywords: Speech Recognition, Arabic Letters, Hybrid Techniques, MFCC, PLP, LPCC, PCA and FFBPNN

1. Introduction

This template, Speech is a basic information vector that facilitates our daily life. It is the fact to translate the meanings in our minds into serial movements of our vocal tract in order to produce interconnected alphabets which form interconnected words that form sentences. Since Speech interests human life, a great part of researches in telecommunication domain has been concentrated during last decades on automatic speech recognition (ASR).

The automatic speech recognition is the fact to aid a computer interpreting a human voice. It consists in extracting oral messages included in speech signal and analyzing their sets of features. The latters will constitute inputs for recognition systems or classifiers which lead to different performances called the system performance.

Nowadays, the automatic speech recognition attracts the attentions of researchers due to the development of communication tools (computers, mobiles, internet, etc.). Compared to other languages such as English, French, Japanese, Spanish, etc., the research in Arabic speech recognition is poor due to the complexity of such language in different levels: Phonetic, linguistic, semantic, contextual, morpho-syntactic, etc.

Moreover, in Arabic language there are three classes: Standard or Classical Arabic (CA) which is the language of the Quran, Modern standard Arabic (MSA) which is used in media and studied in schools, and finally the dialect which is the spoken language that varies from one country to another or even from one region to other in the same country.

Our approach is motivated by this complexity and the lack of researches available in Arabic language speech recognition, that's why in this paper we focus on Arabic alphabet letters the joining of which generates words and consequently sentences.

A phoneme is a minimal unit that serves to distinguish between meanings in words.

In Arabic, there are 34 phonemes six of them are vowels and 28 are consonants [28]. We distinguish two classes of phonemes: pharyngeal and emphatic which characterize the Semitic language such as Arabic and Hebrew [15].

Arabic alphabets are used in many other languages beside Arabic such as Persian and Urdu. The allowed syllables in Arabic are: CV, CVC, and CVCC [21]. C represents a consonant and V represents a long or a short vowel [28]. In spoken Arabic, consonants are followed by four short vowels: "fatha": it represents the /a/ sound and is an oblique dash over a letter, "dhamma": it represents the /u/ sound and has the shape of a comma over a letter, "kasra": it represents the /i/ sound and is an oblique dash under a letter and "soukoon"

which has the shape of a little circle over a letter [28].

Besides, these characteristics that make the recognition hard, the speech signal has a property to be non-stationary.

A normal speaker never pronounces the same alphabet two times identically because the speed and the period of uttering can vary from one time to another. Moreover, when the vocal tract is altered, the speech signal changes, the inter-speaker variability is evident, the same thing for the pitch, intonation and accent that vary with sexes, social, regional and national origins [7, 9, 12, 18].

The paper is organized as follows:

In section 2, we present some related works while in section 3 we describe different features extraction methods. In section 4, we present our proposed speech recognition system. Finally, the section 5 illustrates the experimental results and interpretation followed by conclusion.

2. Relates Works

Various methods and subjects were treated in Arabic speech recognition. Some researchers were interested in speaker identification with mono-speaker or multi-speaker recognition, independent or dependent speaker recognition. Some others were concerned by isolated words or continuous speech.

Due to advanced techniques, Speech recognition becomes an active research area. Satori H. and al. [24] have proposed a spoken Arabic recognition system, where Arabic alphabets were investigated to form the ten Arabic digits (from zero to nine). The proposed system consists of two steps:

- Mel Frequency Cepstral Coefficients (MFCC) features extraction;
- Classification and recognition conducted by CMU Sphinx4 which is a speaker independent system based on hidden Markov model.

The mean performance results reached, when realizing three tests, were between 83.33% and 96.67%.

Al Azzawi Kh. and al. [2] have proposed a hybrid method for automatic recognition of Arabic vowels. Feature extraction was realized by Wavelet Transform (WT) with Linear Prediction Coding (LPC). In the classification phase Feed-Forward Back Propagation Neural Network (FFBPNN) is used which performance obtained was 82.47%.

El-Mashed Sh. Y. and al. [14] have been interested in their paper on connected Arabic digits (numbers) where speaker independent 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, MFCC features extraction of these digits, application of K-means clustering algorithm for the latter features in order to extract the relevant information and finally Support Vector Machine (SVM) is used where it yields to 94% accuracy.

Another work on the recognition of Arabic alphabets based on telephony Arabic corpus is realized by [5], these alphabets were recognized from a corpus developed by King Abdulaziz City for Science and Technology (KACST). The system is based on Hidden Markov Model (HMM) strategy carried out by Hidden Markov Toolkit (HTK); the performance obtained was 64.06%.

In [3], 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.

Arabic speech recognition (ASR) has attracted also the attention of [6] that introduced a genetic algorithm for Arabic handwriting character recognition and then Hopfield artificial neural network is applied. The recognition is divided in four phases: segmentation of the word into characters, pre-processing of each character, extraction and selection of character features and then word recognition. This research reached promising results with accuracies 99%; 92.13% and 90.52% respectively for the training, the validation and the testing sets.

A statically analysis of Arabic phonemes for continuous Arabic speech recognition using a widely used Arabic corpus is a work realized by [22] based on the (HMM). He showed that phonemes, which are based on statistical information, can be clustered in groups. An Arabic alpha digit recognizer was established by [13] where three subsets were used. When using a digit subset, the system recognized Arabic digits with 94.13% accuracy. In the case of alpha subset, alpha recognition is 64.06%, but when mixing alphas and digits subsets the recognition jumps to 76.06%.

Previously, Ganoun A. and al. [17] 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. In [8], 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).

Recently, Daqrouq K. and al. [11] 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%, also the recognition performance in noisy environment has been investigated and the obtained results were very promising.

3. Theoretical Background

To realize our system, two phases are required:

3.1. Parameterization Phase

3.1.1. Cepstral Coefficients

The speech signal varies permanently in time according to the movement of the vocal tract; consequently analysis must be processed on short slide overlapped windows as a speech signal which is considered to be stationary in a short time interval. The speech signal is the result of the convolution in time domain of the source and the vocal tract (filter) [23]:

$$s(n) = e(n) \times h(n) \tag{1}$$

Where s(n) is the filter output, e(n) is the excitation signal and h(n) is the impulse response of the filter. In order to replace the convolution by an addition operation, one passes to the log-spectral domain by the following equation:

$$\log(S(f)) = \log(E(f)) + \log(H(f)) \tag{2}$$

Where S(f), E(f) and H(f) are the Fourier transform of s(n), e(n) and h(n) respectively.

The real Cepster of the speech signal is obtained by applying the inverse of discrete Fourier transform (IDFT) to equation (2), then separation of the source (excitation signal) and the vocal tract (Transfert function) is realized by a time windowing called 'Liftrage' resulting to Cepsral Coefficients. This stage is also called 'homomorphic analyses and it's widely spread in automatic speech recognition domain.

3.1.2. Mel Frequency Cepstral Coefficients (MFCC)

The MFCC are computed by discrete cosine transform of the power spectrum of the speech signal. It is based on the Mel scale which models the perception of the speech by the human ear.

The Mel scale behaves linearly between zero and 1000 Hz and logarithmically above, so it spaces small values and approaches large values: the main advantage of MFCC coefficients is that they are uncorrelated.

To extract MFCC coefficients five steps are employed:

The first step is to pre-emphasis the speech signal by applying a high pass filter in order to increase the high frequency contribution. In fact, when spreading via air, the magnitude of speech signal reduces as the frequency rises. In order to compensate the attenuated speech signal, it is passed through a high-pass filter (finite impulse filter) to recover the signal.

In practice, we use simply a finite impulsion filter (1,-0.97).

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

$$s_p(n) = s(n) - 0.97s(n-1) \tag{3}$$

The second step is to window the speech signal by overlapped Hamming windows. These windows are of little sizes (about 25 ms) and are used to reduce the discontinuity and to avoid the leakage effect and consequently to improve the analysis of the speech. The Hamming window is given by:

$$h_1(n) = \begin{cases} 0.54 - 0.46 \times \cos(2\pi n / N - 1) \; if \; 0 \le n \le N \\ 0 \; otherwise \end{cases} \tag{4}$$

Where N is the size of the window.

This window was chosen since it generates lesser oscillations than other windows and has reasonable side lobe and main lobe characteristics which are required for the DFT computation. The hamming window has effectively better selectivity for large signals and is commonly used is speech processing.

The third one is to compute the Discrete Fourier Transform of each windowed frame resulting in Short Time Discrete Fourier Transform (STDFT). The values derived from here are then grouped together in critical bands and weighted by a triangular filter bank counting M filters called 'Mel-Spaced filter bank'.

The Mel scale is given by the following equation:

$$f_{Mel} = 2595 \times \log(1 + f_{HZ}/700) \tag{5}$$

Where f_{Hz} is the frequency in Hz.

In the fourth step, the logarithm of the band passed frequency response is computed. Finally, the Discrete Cosine Transform (DCT) is applied on the found data which results in Mel Frequency Cepstral Coefficients [27].

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:

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

The Mel Frequency Cepstral Coefficients of the filtered information by the m^{th} filter are represented by $c(m)$ as:

$$c(m) = DCT(\log(X_m)) \tag{7}$$

3.1.3. Linear Predictive Cepstral Coefficients: LPCC

The LPCC feature extraction is based on the LPC analysis which computes the Linear Predictive Coefficients, so the LPCC are calculated from the autoregressive modeling of the speech signal. They are very simple and well used since they allow a good representation of speech overlap vowels.

Each frame is represented by static coefficients: In our work thirteen or sixteen coefficients are used in one hand. In the other hand we have increased the dimensionality of the MLP input vector which represents a letter by concatenating two or more vectors in order to form one vector to improve the recognition task.

After pre-emphasizing and windowing the signal, the autocorrelation features are extracted then the Levinson Durbin is used for computing linear predictive coefficients (LPC) since the vocal tract is modeled by a digital all-pole filter. Finally the linear predictive Cepstral Coefficients (LPCC) for a speech frame are calculated by using the following formula:

$$\hat{v}[n] = \ln(G) \quad for \quad n = 0 \tag{8}$$

Where G is the gain of the all-pole filter (the vocal tract);

$$\hat{v}[n] = a_n + \sum_{k=1}^{n-1} (k/n)\ \hat{v}[k] a_{n-k} \tag{9}$$

for $1 \le n \le p$ [10].

Where $\hat{v}[n]$ is the n^{th} linear predictive Cepstral Coefficient; p is the order of the LPC desirable analysis and a_n is the n^{th} linear predictive coding coefficient computed with the Levinson- Durbin algorithm.

3.1.4. The Perceptually Based Linear Prediction Analysis: PLP and Rasta-PLP

The PLP technique uses several operations inspired of perceptual data: that's to produce a hearing spectrum with the integration of few critical bands in the Bark scale, taking into account the isotone curve, compression of the spectrum in sound intensity and it is based on that of the LPCC.

We just add three steps such as:

- Integration of critical bands;
- Equal loudness pre-emphasis;
- Intensity-loudness conversion to simulate the power low of hearing.

The Rasta-PLP is based on the PLP method. It applies a regressive filter for analyzing and reducing noise [20].

Rasta-PLP is performed in few steps as shown in Figure 1.

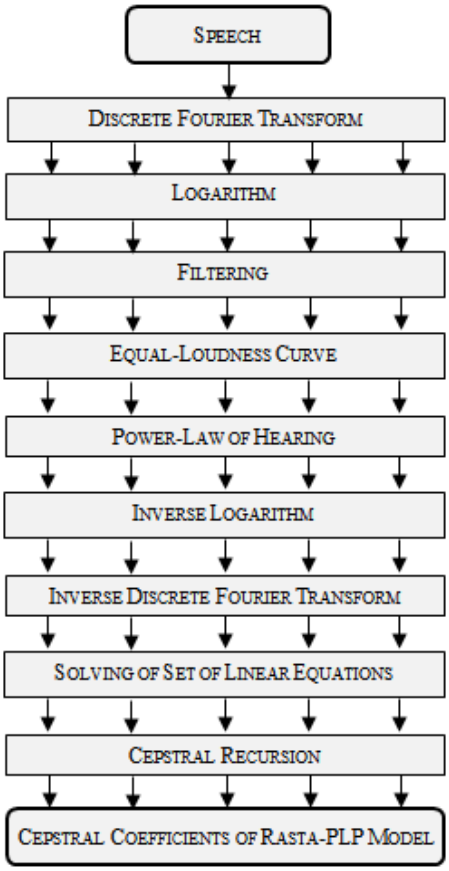

Figure 1. *Steps of Rasta-PLP*

- Compute the critical-band spectrum (as in the PLP) and take its logarithm;
- Apply a regressive filter for analyzing and reducing noise ;
- In accord with the conventional PLP, add the equal loudness curve and multiply by 0.33 to simulate the power low of hearing;
- Take the inverse logarithm of this relative log spectrum, yielding a relative auditory spectrum;
- Compute an all-pole model of this spectrum, following the conventional PLP technique [16].

3.1.5. Principal Component Analysis (PCA)

The principal component analysis technique is used for reducing dimensionality of the obtained features by conserving the intrinsic original information. We used PCA as a modeling tool of the extracted features because it is a simple, non-parametric method of extracting relevant information from confusing data sets. Here our purpose of using PCA is to facilitate the recognition since this technique allows us to represent each letter by a minimum number of vectors [25]. In practice and to apply PCA, we follow the steps below:

- Calculate the covariance matrix of the features on which we will apply PCA;
- Find the eigenvectors of the obtained covariance matrix;
- Extract diagonal of matrix as vector;
- Sort the variances in decreasing order;
- Project the original data set.

3.2. Recognition Phase

In the recognition phase, we have used the Multilayer Perceptron (MLP) Feed-Forward Back Propagation Neural Networks which is known as the most popular Multilayer architecture. It is formed by an input layer (Xi), one intermediary or hidden layer (HL) and an output layer (Y). A weight matrix (W) can be defined for each of these layers.

This artificial neuronal network topology can solve classification problems involving non-linearly separable patterns and can be used as a universal function generator [19]. Important issues in MLP design include specification of the number of hidden layers and the number of units in these layers. The number of input and output units is defined by the problem (there may be some uncertainty about precisely which inputs to use) [26].

In our work, the neural network has been trained in supervised mode, we used a binary code of 7 bits as a Target, we choose a number of neurons between 50 to 130 and the "TanSig" activation function for the hidden layer. For the output layer, we choose seven neurons and the "LogSig" activation function, the learning algorithm was stochastic gradient descent, the used epochs have been varied between 17 and 500. The performance function is mean square error (MSE) and the training function is 'Trainlm'. The remaining parameters are taken by default.

4. The Proposed Speech Recognition System

The proposed speech recognition system consists of three modules according to their functionalities.

The first module concerns the recording phase which is followed by an enhancement procedure in order to obtain a best and intelligible quality of signal.

In the second module features are extracted and transformed on reduced data variables keeping the most discriminatory information.

The last module is used in the recognizing phase which includes training and testing processes.

The used corpus of experimentation is based on all Arabic letters with their four vowels pronounced by four speakers. The main features were extracted by using separately different known techniques in speech analyzing domain (MFCC, PLP, etc.) and transformed to a reduced data with the principal component analysis (PCA) procedure. The obtained data is used to train a neural network (Feed Forward Back Propagation Neural Network). In this work, we have changed the feature extraction techniques applied in [1]. In fact, several hybrid techniques have been experienced such as MFCC combined to PCA, PLP followed by continuous wavelet transform (CWT) and PCA. In this work, the corpus of test is composed of sequences completely different from the training corpus.

4.1. Corpus Preparation

The corpus is interested in automatic recognition of Arabic letters. Four speakers (two males and two females) were participated to build it by uttering all Arabic letters (28) with their four vowels three times each one firstly and secondly five times.

For the first case (three trials): the number of utterances of each speaker is equal to: $28 \times 4 \times 3 = 336$ utterances, then the total number of utterances for the four speakers was equal to 1344. Each utterance is put in a separate wave file and each trial of each speaker is put in a separate sub corpus.

Suppose for example that Sal, AFi, Fat and Sou are the four speakers, so Sal corpus is composed of three sub corpora of 112 utterances each.

The training corpus is composed of the two first trials of the four speakers (Sal_1, Sal_2, Afi_1, Afi_2, Fat_1, Fat_2, Sou_1, Sou_2), so the number of files used here is equal to 896 files.

The validation corpus is composed of the second trial of Afi (Afi_2 =112files) and the first trial of Fat (Fat_1=112files). So the total number of files in validation corpus is equal to 224 files.

The test corpus is composed of the third trial of Sal (Sal_3) and the third trial of Sou (Sou_3). So the total number of files in test corpus is equal to 224 files.

Thereby, the train corpus is composed of 80% of the total corpus and we choose to construct the test corpus by 20% of the total corpus from who hasn't been included in the train corpus. The validation corpus is composed of 20% of the total corpus from that has been included in the train corpus.

4.2. Data Acquisition

The recording of the speech has been occurred in a suitable environment, with professional materials in a professional acoustic studio: A digital mixing console (Studer 2000M2) and a dynamic microphone MD421. The speech signal of each letter is recorded and digitized with a sound card of a computer equipped by "Sound Forge" software. The parameters of the recordings were: Mono wave files, a sampling rate of 44100 Hz and a 16 bits resolution. During all the recording sessions, each utterance was played back to ensure that the entire signal of each letter was recorded then stored as a Mono wave file in a corresponding sub corpus.

4.3. Feature Extraction

For all the whole work, feature extraction and recognition were implemented in Matlab7.1 platform language. Each speech signal corresponding to any letter is put in a specific file. The stages of analysis have been occurred in the following steps.

- Extract each file corresponding to each letter from its sub corpus and read it by using the corresponding Matlab command;
- Remove Silence and reduce noise in the signal obtained in the first step in order to improve quality and enhance the speech signal by applying the algorithm of "Minimum Mean Square Error Short Time Spectral Amplitude Estimator" (MMSE-STSA);
- Apply one of the feature extraction techniques which have been already mentioned above to the signal obtained in step 2;
- Apply PCA to the extracted features resulting from step3, in order to represent the letter with the minimum number of vectors;
- Concatenate the obtained vectors in order to obtain one vector which will represent the speech signal for one letter. Our purpose from concatenating these vectors is to simplify computing and improve recognition performance.
- Do the same steps for all Arabic letters;
- Put all the final total vectors (these represent the total letters) obtained in step5 in one matrix. Each column in this matrix represents one letter;
- Select from the matrix obtained in step7 the number of vectors which are designed to build the different corpora (training, validation and test) taking into account the size of each one. On the one hand, we have chosen 80% of vectors obtained in the latter matrix to construct the train corpus. On the other hand, 20% are selected for the test corpus and 20% among that of the train corpus to build the validation corpus.

Each corpus will be put in a matrix and constitute an input for the feed-forward back propagation neural networks.

5. Experimental Results and Discussion

The features extraction techniques mentioned above (such

as MFCC, PLP, etc.) have been applied to speech signal corresponding to each letter to provide Matrices. The PCA technique is then applied to these matrices to reduce the dimensions. In order to represent each letter with one vector, vectors of each Matrix were concatenated. All obtained vectors which represent the total letters are grouped in one matrix. The final matrix is provided to Multilayer Perceptron MLP as inputs. The number of neurons in the hidden layer has been varied between 50 and 130, the goal was G=0.01. We let the Matlab program prepared for our recognition system running until one of the known MLP stop criterions is reached, and we note each time the corresponding error rates.

Another independent stage has been done when we conserved the same parameters already used and we increased the trial number. Instead of uttering three times each letter, each speaker is invited to utter each letter five times. This latter operation has significantly improved the recognition performance.

It is found that when we concatenate feature vectors which represent the speech signal of any letter in one vector, we obtain a better result for speech recognition. In addition, the choice of the number of vectors to concatenate and the convenient parameters for the neural network are very interesting. The extension of the trials number of uttering letters from three to five times and consequently the extension of speech corpus has improved the recognition performance with all feature extraction techniques.

Compared to all used feature extraction methods tested in this work, the Perceptual Linear Prediction Technique (PLP) occupies the first order in term of recognition performance and in term of computing time.

After different experiments, we have reached the following error rates as presented in tables I and II.

Table II shows that performance is far better improved when the number of trials or recordings per person was increased.

The best performance was obtained by using the PLP technique combined with PCA. This outperformance can be interpreted by the fact that 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.

The same feature vectors (PLP and PCA), already used, were also computed as inputs for RBF neural networks. The obtained results show that RBF neural networks respond poorly when using large training vectors. The reached performances were respectively 587.056e-030%, 261.71e-030 % and 27.98% for training, validation and test. These results prove that FFBPNN is more efficient than RBF neural networks.

Finally, the use of several hybrid techniques for feature extraction phase, the concatenation of input vectors and the Feed-Forward Back Propagation Neural Network has significantly improved the recognition systems. However, in order to obtain more efficient results, the experimental corpus should necessary be extended in terms of speaker numbers and different Arabic dialects.

Table 1. *Recognition performance by using three trials per person*

Method	Training Error in %	Validation Error in %	Testing Error in %	# of Epochs
MFCC + PCA	1.187189	0.897189	18.3939	930
ΔΔMFCC + PCA	0.991074	0.860687	30.4044	177
MFCC + ΔMFCC + ΔΔMFCC + PCA	8.29687	9.33385	23.5392	232
PLP + PCA	0.978391	1.20843	12.6183	27
Rasta PLP + PCA	0.732753	0.715146	17.5906	17
LPCC + PCA	0.99792	1.02594	23.8582	83
CWT + PCA	3.30559	3.09105	26.1099	490
CWT + PLP + PCA	0.961214	0.783404	24.8872	36
PLP + CWT + PCA	7.48001	7.10607	21.6927	72
CWT + MFCC + PCA	0.987389	0.837862	27.5388	41
MFCC + CWT + PCA	4.48791	4.95848	23.2967	146
CWT + LPCC + PCA	1.40306	1.21173	32.2703	91

Table 2. *Recognition performance with five trials per person*

Method	Training Error in %	Validation Error in %	Testing Error in %	# of Epochs
PLP + PCA	0.992308	0.814672	*11.7432*	179
MFCC + PCA	0.890573	0.664166	13.1609	53
Rasta PLP + PCA	0.936744	0.819379	16.9822	30
LPCC+PCA	1.73818	1.72194	18.4507	271
MFCC + ΔMFCC + ΔΔMFCC + PCA	6.93287	6.8629	20.0726	78

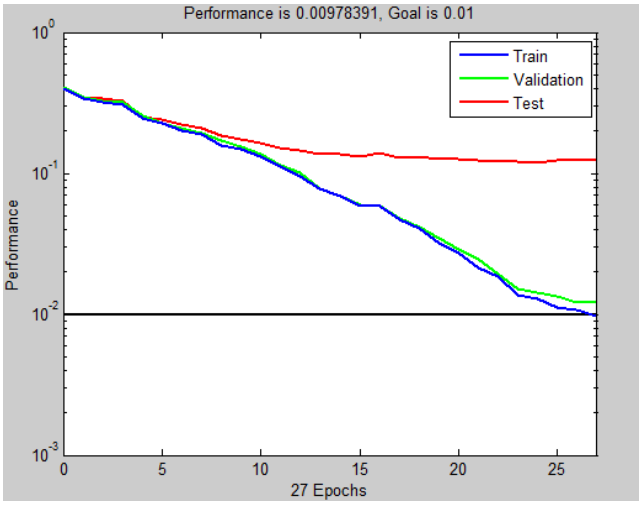

Figure 2. *Performance curves by using three trials with PLP and FFBPNN techniques.*

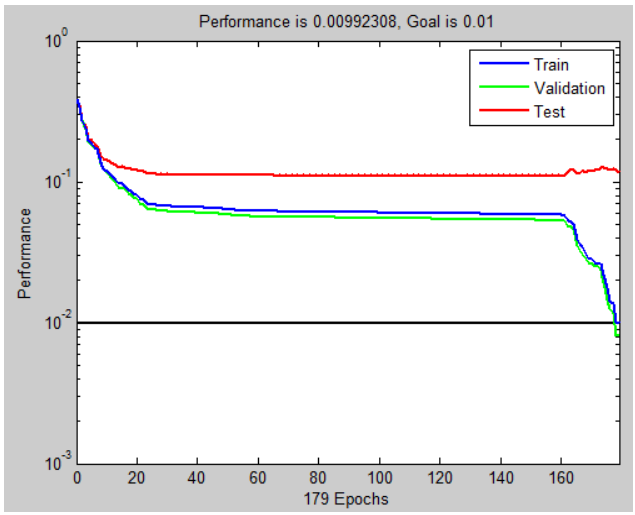

Figure 3. *Performance recognition curves by using five trials with PLP and FFBPNN techniques*

6. Conclusion

In this paper, a variety of feature extraction techniques and a feed forward back-propagation neural network (FFBPNN) have been tested for automatic speech recognition system of Arabic letters. For this purpose, a proper corpus was prepared involving recordings of four speakers.

In the first set of experiments, each speaker was invited to utter the total letters three times. Obtained data were performed with a variety of approaches and hybrid techniques. In this case PLP technique associated to PCA algorithm have presented the best performance with a testing error equals to 12.6183%.

In the second set of experiments, the corpus was improved by increasing the number of trials of each speaker to five. In this second case, recognition performances were far better with all used approaches. However, the PLP technique still the best one and presented a testing error of 11.7432%.

Compared to RBF neural networks, FFBPNN has given more satisfactory results in terms of training, validation and test performances.

In the future work, we plan to extend the experimental corpus and to develop a new system based on some other advanced techniques such as Bionic Wavelet Transform (BWT) for the feature extraction and Support Vector Machines (SVM) for the recognition phase. This research can be used in order to improve isolated word recognition and Arabic dialect recognition.

References

[1] Abdulfattah Ahmad M. and El Awady R. M., "Phonetic Recognition of Arabic Alphabet Letters Using Neural Networks," International Journal of Electric & Computer Sciences IJECS-IJENS, Vol. 11, No. 01, 112501-3434 IJECS-IJENS ©, February 2011.

[2] Al Azzawi Kh. Y. and Daqrouq Kh., "Feedforward Backpropagation Neural Network Method for Arabic Vowel Recognition Based on Wavlet Linear Prediction Coding," International Journal of Advances in Engeneering & Technologiy", Ijalet ISSN:2231-1963, Sept. 2011.

[3] Al-Irhaim Y. F. and Saeed E. Gh., "Arabic Word Recognition Using Wavelet Neural Network," Third Science Conference in Information Technology, November 2010.

[4] Alkhouli M., "Alaswaat Alaghawaiyah," Daar Alfalah, Jordan, 1990.

[5] Alotaibi Y. A., Alghamdi M. and Alotaiby F., Computer Engineering Department, King Saud University, Riyadh, A. Elmoataz et al. (Eds.): ICISP 2010, LNCS 6134, pp. 122–129, 2010. © Springer-Verlag Berlin Heidelberg, 2010.

[6] Al-zoubaidy L. M., "Efficient Genetic Algorithm for Arabic Handwritten Characters Recognition," Raf. J. of comp. & Math's, vol.6, No.2, 2009, received on:29/4/2008,Accepted on :3/9/2008.

[7] Barras C., "Reconnaissance de la Parole Continue : Adaptation du Locuteur et Contrôle Temporel dans les Modèles de Markov Cachés," Thesis, Université de Paris IV, 1996.

[8] 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, Vol. 4, No 3, March 2012.

[9] Boite R., Kunt M., "Traitement de la parole," Presse polytechnique romandes, 1987.

[10] Cheng O., Abdulla W. and Sacic Z., "Performance Evaluation of Front-end Processing for Speech Recognition Systems," Electrical and computer Engineering Department School of Engineering, University of Auckland, School of Engineering Report No.621, 2005.

[11] 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 & Technology, 1238-1255, 2014.

[12] Deroo O., " Modèles Dépendants du contexte et Méthodes de Fusion de données Appliquées à la reconnaissance de la Parole par Modèles Hybrides HMM/MPL ", Thesis, Faculté Polytechnique de Mons, 1998.

[13] El-Ghazi A., Daoui C. and Idrissi N. "Automatic Speech Recognition System Concerning the Maroccan Dialecte (Darija and Tamazight)," International Journal of Engineering Science and Technology (IJEST), ISSN: 0975-5462 Vol. 4 No.03 March 2012.

[14] EL-Mashed Sh. Y., Sharway M. I., Zayed H. H., "Speaker Independent Arabic Speech Recognition Using Support Vector Machine," ICI-11 Conference and Exhibition on Information technology and Instruction technology, Hungary 2011.

[15] Elshafei M. "Toward an Arabic Text-to-Speech System," vol. 4B no. 16, pp 565–583, Octobre 1991.

[16] Furui S., "Speaker-Independent Isolated Word Recognition Based on Enphasized Spectral Dynamics," Procs. IEEE Intl. Conf. on Acoustic, Speech & Signal Processing, pp.1991-1994, Tokyo, Japan 1986.

[17] Ganoun A. and Almerhag I. "Performance Analysis of Spoken Arabic Digits Recognition Techniques," Journal of Electronic Science and Technology, Vol. 10, No. 2, June 2012.

[18] Génin J., "La parole et son traitement automatique Calliope," Annales des Télécommunications, vol. 45, Issue 7-8, pp 457-458, August 1990.

[19] Haykin S., "Neural Networks and Learning Machines", Prentice Hall, USA, 2009.

[20] Hermansky H., Morgan N., Bayya A. and Kohn Ph., "Rasta-PLP Speech Analysis", TR-91-069, Decembre 1991.

[21] Kouloughli D. E., "Sur la Structure Interne des Syllabes «lourdes» en Arabe Classique," vol. 16, numéro 1, pp 129-154, 1986.

[22] Nahar K. M.O, Elshafei M., Al-Khatib W. G. and Al-Muhtaseb H., "Statistical Analysis of Arabic Phonemes for Continuous Speech Recognition," International Journal of Computer and Information Technology, ISSN: 2279 – 0764 Vol. 01, Issue 02, November 2012.

[23] Rabine L. and Schafer, R., "Digital Processing of Speech signals", Prentice Hall, 1978.

[24] Satori H., Hiyassat H., Harti M. and Chenfour N., "Investigation Arabic Speech Recognition Using CMU Sphinx System," The International Arab Journal of Information Technology, Vol. 6, No. 2, April 2009.

[25] Shlens J., "A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS," Derivation, Discussion and Singular Value Decomposition, March 2003.

[26] Venkateswarlu R.L.K., Kumari R. V. and Vani Jayasri G., "Speech Recognition Using Radial Basis Function Neural Network", IEEE, 2011.

[27] Zabidi A., Mansor W., Khuan L. Y., Sahak R. and Rahman F. Y. A., "Mel-Frequency Cepstrum Coefficient Analysis of Infant Cry with Hypothyroidism," 5th Int. Colloquium on Signal Processing & Its Applications, Kuala Lumpur, Malaysia, 2009.

[28] Zitouni I., Sarikaya R., "Arabic Diacritic Restoration Approach Based on Maximum Entropy Models," Computer Speech and Language, vol. 23 pp 257–276, july 2009.