World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:7, No:7, 2013

Echo State Networks for Arabic Phoneme Recognition

Nadia Hmad and Tony Allen

***Abstract*—**This paper presents an ESN-based Arabic phoneme recognition system trained with supervised, forced and combined supervised/forced supervised learning algorithms. Mel-Frequency Cepstrum Coefficients (MFCCs) and Linear Predictive Code (LPC) techniques are used and compared as the input feature extraction technique. The system is evaluated using 6 speakers from the King Abdulaziz Arabic Phonetics Database (KAPD) for Saudi Arabia dialectic and 34 speakers from the Center for Spoken Language Understanding (CSLU2002) database of speakers with different dialectics from 12 Arabic countries. Results for the KAPD and CSLU2002 Arabic databases show phoneme recognition performances of 72.31% and 38.20% respectively.

***Keywords*—**Arabic phonemes recognition, echo state networks (ESNs), neural networks (NNs), supervised learning.

I. Introduction

MANY different types of Neural Network (NN) architecture have been used for a variety of speech recognition purposes. These include: Multi-Layer Perceptron (MLP) [1]-[3], Recurrent Neural Networks (RNNs) [4] and Echo State Networks (ESNs) [5], [6].

ESNs are particularly well suited for classifying temporal signals as part of a dynamic pattern recognition task [7]. An ESN with delay and sum readout has been used to perform as a nonlinear Audio signal identification system [8], for time series modeling [9] and as a 10th order NARMA (The Non-linear Auto-Regressive Moving Average) system [10], [11]. Moreover, ESNs have shown good ability for learning grammar structure as part of a natural language task [12], [13]. ESNs, in addition, have been used for pattern recognition tasks. It has been shown that ESNs have the capability to extract low dimensional features from a dynamic reservoir for a handwriting recognition task. The main use of the ESN in this process was to generate the features of data in a high dimensional representation [13]. For complex speech recognition problems, ESN have been used to successfully classify ten isolated English digit [5], [11], [14] and perform Japanese Vowels classification [15].

Recently, an English vowel classification system was investigated in [16]. The AEV vowel database was used to train an Echo State Network (ESN) using a forced supervised learning algorithm. This dataset contains 12 vowels uttered by 48 women, 45 men, and 46 children. Cochlear filtered audio was used as input to the 74 input neurons connected to the network reservoir. The overall performance of the vowel classification system was 81.7% with a maximum performance of 84.2% for male speakers.

The work in [15] investigated the ability of an Echo State Network (ESN) for Japanese Vowel classification. In actual fact, the Japanese Vowels dataset only contained recorded utterances of the one vowel /ae/ for nine male speakers. However, the training and test datasets consisted of 270 and 370 samples respectively. These datasets were used to train ESNs containing different reservoir network sizes ranging from one to 1000 nodes with spectral radius from 0.993 to 0.996. The best performance was 100% classification for network sizes between 500 and 1000.

In [7], an ESN-based speech recognition system was used to predict the English digits. In this system, each frame-based prediction model was used to predict one digit. Human factor cepstral coefficients (HFCC) were used as the feature extraction technique within 20 ms Hamming windows. 12 HFCC coefficients were extracted and used as input to the ESN classifier. The total dataset consist of 4130 utterances of isolated English digits from Zero to Nine from8 female and 8 male speakers. These were split in half to produce a training dataset consisted of utterances from 4 female and 4 male speakers, and a testing dataset contained the same number of speakers. The claimed classification accuracy was 100% and 99.1% for training and test sets respectively with a reservoir network size of 60. In comparison, the same datasets gave 94.7% testing classification accuracy with a conventional Hidden Markov model (HMM).

In this paper ESNs trained with supervised, forced supervised and combined supervised/forced supervised learning algorithms are compared for Arabic phonemes classification and recognition. Section two presents details of the supervised and forced supervised Echo State Networks (ESN). The proposed Arabic databases and feature extraction techniques are then briefly described in section three. Finally, the results of the experiments performed on the proposed Arabic datasets are discussed in section four.

II. Echo State Networks (ESN)

There are several types of training algorithm that can be used to train recurrent networks: real-time recurrent learning (RTRL), back-propagation revisited, back-propagation through time (BPTT), Extended Kalman Filtering techniques (EKF) [17], and Hessian-Free (HF) Optimization [18]. Usually these algorithms result in suboptimal solutions with slow convergence. Echo state networks (ESNs), invented by Jaeger [17], are a novel structure of recurrent neural networks (RNNs) that contain a large, random, fixed and untrained

Nadia Hmad is with the School of Science & Technology, Nottingham Trent University, NG1 4BU, UK (e-mail: N0239625@ntu.ac.uk).

Tony Allen is with Centre for Innovation & Technology Exploitation Schoolof Science & Technology, Nottingham Trent University, NG1 4BU, UK (e-mail: Tony.allen@ntu.ac.uk).

recurrent "dynamic reservoir" network. The learning algorithm of the ESNs is very simple and linear, in that only the weights from the reservoir to the outputs neurons are adapted (see Fig. 1). As a result, the learning process is fast with less computation. The optimal output weights for the ESN are obtained when the *MSE* is minimized.

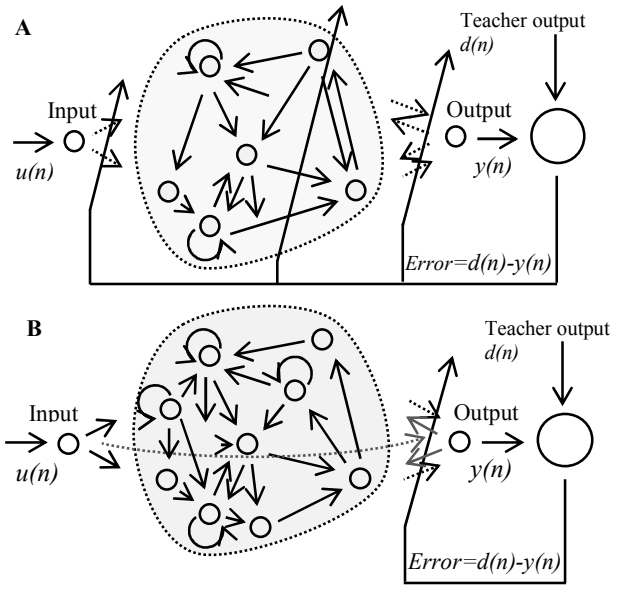

Fig. 1 Comparison between RNNs and ESN architecture and training In A, all the RNN weights are changed during training whilst in B only the ESN output weights are changed [17]

A. ESN Architecture

The reservoir size effectively determines the ESN's performance. In addition, connectivity, spectral radius, reservoir and output activation functions, shift and scale parameters also have effect. The architecture of the ESN used in this project was based on the work of Ted et al. [16] and Jaeger [17]. The ESN connections used are shown in Figs. 2 and 3. During the ESN training and testing, the activations of their servoir and output neurons were calculated using either a supervised or forced supervised learning algorithm.

1. ESN with Supervised Learning Algorithm

Learning systems observe a training dataset constructed from features of instances with its labels as pairs, represented by $\{(x_1, d_1), ..., (x_n, d_n)\}$. The aim of this learning is to predict the output y for any given input feature x.

The activations of the ESN reservoir and output neurons in a supervised learning mode are calculated using (1) and (2):

$$x(n+1) = \tanh(W^{in}u(n+1) + Wx(n) + W^{back}y(n)) + leftover \quad (1)$$

where $x(n+1)$ is the reservoir state for time step $(n+1)$, $u(n+1)$ is the input vector, $y(n)$ is the calculated output for the supervised learning for time step (n), and the *leftover*, as used in [16], is calculated as follows:

$$leftover = x(n) * (1 - d\_leakRate) \quad (2)$$

The calculated output *y(n+1)* is given by:

$$y(n+1) = \tanh(W^{out} * x(n+1) + W^{out} * u(n+1)) \quad (3)$$

where $y(n+1)$ is the calculated output states for time step $(n+1)$.

Weight vectors W^{in}, W, W^{back}, and W^{out} are initially generated with random values between -1 and 1, and with a connectivity parameter between 0 and 1 in order to generate random connections between neurons (see Fig. 2).

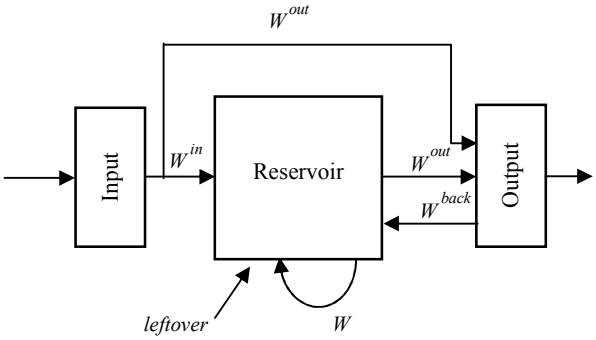

Fig. 2 ESN connections with supervised learning, the weight connections in the figure as follows: W^{in} is input weight vector, W is reservoir weight vector, W^{out} is the output weight vector, W^{back} is the feedback weight vector, and *leftover* is a variable calculated using (2)

The supervised ESN network was implemented and optimized for the KAPD and the CSLU2002 Arabic speech datasets. The final network parameters are shown in Table I.

2. ESN for Forced Supervised Learning

Forced supervised learning replaces the actual output $y(n)$ by a teacher output (*desired output*) $d(n)$ during the training stage [16]. This is often called *Teacher-Forced* supervised learning.

For the forced supervised learning ESN, the activations of reservoir and output neurons are calculated using (4) and (5):

$$x(n+1) = \tanh(W^{in}u(n+1) + Wx(n) + W^{back}d(n)) + leftover \quad (4)$$

where $x(n+1)$ is the reservoir state for time step $(n+1)$, $u(n+1)$ is the input vector, $d(n)$ is the desired output vector (target output vector) for time step (n), and the *leftover* was calculated using (2).

TABLE I
SUPERVISED ESN PARAMETERS

Parameter	Value
Reservoir size:	400
Connectivity:	0.2
Spectral Radius:	0.996
Activation function:	tanh
Input size:	11
Input connectivity:	0.5
Input Shift:	0
Input Scale:	1
Output size:	33
Output activation:	tanh
Feedback connectivity:	0.3
Feedback Shift:	0
Feedback Scale:	1
D_leakRate	0.0
Alfa	0.33
Wash-out time	0

$$\hat{y}(n+1) = \tanh(W^{out} * x(n+1) + W^{out} * u(n+1)) \quad (5)$$

where $\hat{y}(n+1)$ is the calculated output states for time $(n+1)$.

Weight vectors W^{in}, W, W^{back}, and W^{out} are randomly generated as described previously using the connections shown in Fig. 3 and main the parameters are as described in Table I.

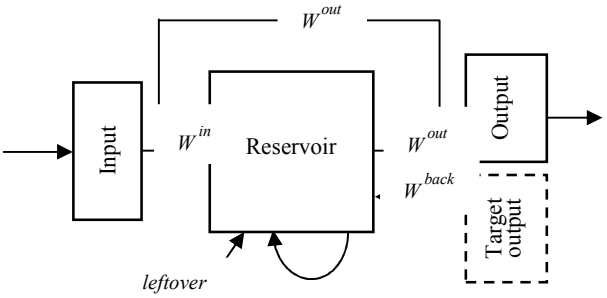

Fig. 3 ESN connections with combined supervised/forced supervised learning algorithm, the weight connections in the figure as follows: W^{in} is input weight vector, W is reservoir weight vector, W^{out} is the output weight vector, W^{back} is the feedback weight vector, and *leftover* is a variable calculated using (2)

3. ESN for Combined Supervised/Forced Supervised Learning

Unfortunately, an ESN trained with the standard forced supervised learning algorithm is not suitable for recognition purposes. This is because the target output would be required for calculating the reservoir states during the testing stage. To address this shortcoming, a combined supervised/forced supervised learning was implemented. In this novel algorithm, the calculated network outputs are first passed through a maximum likelihood stage to effectively produce teacher outputs before being fed back into the reservoir (see Fig. 4). The target outputs used as feedback connections are binary values zero or one, whilst the calculated outputs are floating point numbers between -1 and 1.

During testing, all the outputs were converted from floating point numbers into zeros or ones using the maximum likelihood algorithm.

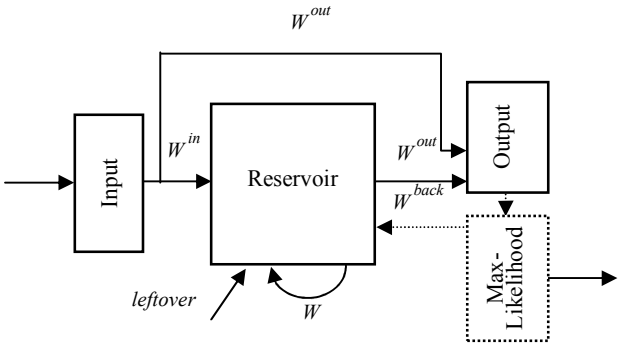

Fig. 4 ESN connections with modified supervised/forced learning algorithm, the weight connections in the figure as follows: W^{in} is input weight vector, W is reservoir weight vector, W^{out} is the output weight vector, W^{back} is the feedback weight vector, and *leftover* is a variable calculated using (2)

4. Updating Weight Vectors

In all ESN architectures presented here, the outputs weights are updated using the offline pseudo inverse method to calculate the new output weights W^{out} after each training epoch. The following steps describe this algorithm:

- An autocorrelation matrix (state matrix) A of size $(n \; X \; (N+K))$ is accumulated from the state vector $x(n)$ and input vector $u(n)$ for each time step n.
- A cross-correlation matrix (output matrix) B of size $(n \; X \; L)$ is also accumulated from the target output $d(n)$ for each time step n.
- The *pseudo inverse matrix* is then calculated using (6)

$$W^{out} = ((A^T A + \alpha^2 I)^{-1} A^T B)^T \quad (6)$$

where I is the identity matrix, $(A)^{-1}$ is the inverse matrix of A, and A^T donates the transpose of matrix A, and α is the smoothing factor where $\alpha \leq 1$. When $\alpha = 0$ (6) reverts to a linear regression (the Wiener-Hopf solution [16]) as in (7):

$$W^{out} = ((A^T A)^{-1} A^T B)^T \quad (7)$$

The strongest regularization is obtained when $\alpha = 1$.

III. DATABASE AND FEATURE EXTRACTION

A. Arabic Databases

The KAPD contains 340 utterances containing each of the 33 Arabic phonemes presented in 12 semi-words spoken by 7 male speakers in a quiet environment with a Saudi Arabia

dialectic. This database recorded using microphone and sampled at 10 kHz 16-bit. This KAPD dataset was previously used for Arabic phoneme classification using a conventional multi-layer perceptron neural network [3].

The CSLU database consists of 22 group datasets of telephone speech for 22 languages. Eastern Arabic is one of these languages. The Arabic corpus contains fluent continuous speech of 98 native callers (5 Childs, 23 females and 70 males) from more than 12 countries. Each speaker in this corpus talks about fixed topics (e.g. address, time home town, weather etc.). A total of 1997 sentences files collected over telephone lines in different environments for the Arabic database and sampled at 8 kHz 8-bit. 34 speakers (17 females and 17 males) were selected arbitrarily from the Arabic dataset. 3802 phonemes were manually segmented from this database to train and test the ESN for speech classification and recognition tasks. This resulted in 1894 phoneme samples for training and 1908 phoneme samples for testing.

B. Challenges of Arabic Databases

The KAPD database is very limited as it consists of only 7 male speakers. In addition, one of the speakers' dataset is also very noisy, meaning that it was excluded from the experiments. This dataset is, however, used here in order to allow direct comparison between the ESN results it provides and our previous work with an MLP based Arabic phoneme recognition system [3].

The CSLU2002 databases as described are recorded using a telephone channel. As a result, the quality of the utterances in the databases is dependent on the speakers' environments with several of the utterances containing car and home noise such as baby crying and children talking. Moreover, the length of the utterances is dependent on the speaker. Several of speakers were talkative, producing many phonemes. Some of recorded speech, however, contained very limited speech, producing recorded utterances that contained few or even zero examples of some phonemes. An additional problem is that speakers were from 22 different countries; consequently, they have different dialectics. Several of these dialectics omit some phonemes. For example, the Egyptian dialectic misses phonemes /q,ق/, /j,ج/, /!,ذ/, /Z,ظ/whilst the Lebanese dialectic misses /!,ذ/ and /Z,ظ/. Furthermore, phoneme pronunciation is completely different from one dialectic to another. For example, phoneme /!,ذ/ in the Egyptian dialectic is pronounced as /z,ز/ - a standard Arabic sound. Likewise, the phoneme /q,ق/ in the Yamane dialectic has a sound that has no similar in standard Arabic sounds. Finally, different speakers had different speed of speech during recording.

C. Feature Extraction Technique

The Mel-Frequency Cepstrum Coefficients (MFCCs) algorithm was used to extract features from the KAPD and the CSLU2002 corpus. A frame window of 6.4 ms over lapped by 50% was used for the speech signal frames. Each frame was represented by 12 cepstral coefficients.

The Linear Predictive Code (LPC) technique was also used to extract features from the KAPD for input feature comparison purposes.

IV. Experiments Results

A. Results Using the KAPD Database

1. Closed Set Speaker Independent ESN with Supervised Learning

For comparison purposes, two KAPD databases were created – one being the raw data and the second being anormalized version of the raw data [3]. Both KAPD Arabic phonemes datasets were used to train the ESN with the supervised learning algorithm as described in Fig. 2 and Table I. The results of normalized and raw databases are shown in Fig. 5.

It is clear from Fig. 5 that, for the raw training dataset, most of KAPD Arabic phonemes are recognized with 90% to 100% accuracy and there are no phonemes recognized below 70%. In comparison most of phonemes in the normalized KAPD training dataset were classified with only 70% to 80% accuracy. Moreover, for the testing dataset, most of the KAPD raw Arabic phonemes were recognized between 70% and 80% accuracy; with just two of the testing phonemes being recognized with performance less than 50%. In contrast, the average recognition of the KAPD normalized Arabic phonemes was between 60% and 70%, and approximately 13 phonemes recognized with performance less than 50%.

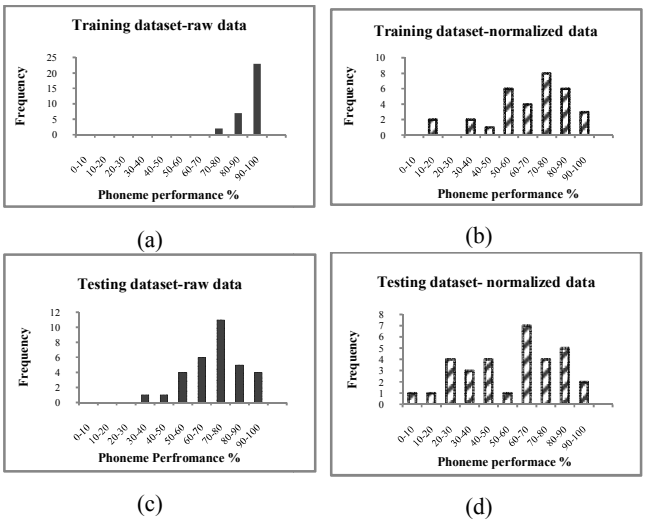

Fig. 5 Histograms for Arabic phonemes recognition performances for training dataset for raw data (a) and normalized data (b), and for testing dataset for raw data (c) and normalized data (d)

TABLE II
A Comparison between the Performance of the ESN and MLP Networks Supervised Learning Algorithm for the KAPD

Neural Network type	Performance
ESN	72.31
MLP	47.36

A comparison between the ESN performance and the performance of a Multi-Layer Perceptron (MLP) used in our previous work [3] was also carried out using the same

database (the KAPD) - see Table II.

In [3], the KAPD Arabic dataset was also categorized into five classes (Stops, Fricatives, Nasals, Letretive and Vowels). For comparison purposes this experiment was also performed on the ESN network with supervised learning algorithm using the same network structure as detailed in Fig. 2 and Table I. A comparison between the average results of Arabic phonemes recognition for the five Arabic categories using the MLP NN and ESN are shown in Table III.

It is obvious from the results that the ESN has a better performance than the MLP NN system for all classes. The results also show that for the ESN system the Stops class (class 1) has the lowest accuracy on the testing dataset. This is probably due to the similarity in phoneme's pronunciation. Stops phonemes are also very short phonemes and hence contain little information with which to classify.

TABLE III
A COMPARISON BETWEEN THE PERFORMANCE OF THE ESN SUPERVISED AND MLP NETWORKS LEARNING ALGORITHM FOR THE KAPD DATABASE

Class	Performance	
	MLP	ESN
Stops	59.90	64.67
MLP fricatives	59.32	81.00
Nasals	77.78	85.57
Letretive	88.35	98.01
Vowels	58.92	79.65

2. MFCC's Features verse LPC Features Using Closed Set Speaker Independent ESN with Supervised Learning

The same KAPD Arabic phonemes dataset was used to compare the performances of the ESN when trained with the supervised learning algorithm on the KAPD dataset using the LPC and MFCC feature extraction techniques: see Table IV.

3. Open Set Speaker Independent ESN with Supervised Learning

The ESN with forced supervised leaning was used for a speaker independent test (open-set). Speakers 6 and 7 in the KAPD were not used during training of the network. During subsequent testing of the trained ESN network, the average phoneme classification accuracy for these two speakers were 40.79% and 39.12% respectively.

TABLE IV
A COMPRESSION BETWEEN THE PERFORMANCE OF THE MFCCS AND LPC FEATURE EXTRACTION TECHNIQUES USING ESN WITH SUPERVISED LEARNING ALGORITHM FOR THE KAPD

Feature extraction technique	Performance %
MFCC	92.92
LPC	66.75

B. Results Using the CSLU2002 Dataset

ESN networks with supervised and combined forced supervised/supervised learning algorithms were used to train and test on the CSLU2002 Arabic database in different experiments. Firstly, each gender of this database was examined separately using the ESN with the different learning algorithms (supervised learning forced supervised learning and combined supervised/forced supervised learning). Secondly, a combined 'female & male' dataset were collected to build a final CSLU2002 Arabic phonemes database. This was then tested using all ESN learning algorithms.

1. Closed Set Speaker Independent ESN for CSLU2002 Female Speakers

17 female speakers were arbitrarily selected from the CSLU dataset and used to create a CSLU2002 female phonemes dataset. This dataset was used to train and test on the ESN with the supervised and forced supervised learning algorithms. Fig. 6 shows the average results of Arabic phonemes for the 17 female speakers.

As you can see from Fig. 6, the ESN trained with the combined supervised/forced supervised training algorithm can correctly classify the majority of the phonemes from CSLU2002 female speakers with an average phoneme recognition performance of 83.84% for the training dataset. However, the average performance of the testing dataset phonemes is 45.78% phoneme recognition. For the ESN trained with the supervised learning algorithm, the average result of phoneme recognition was 73.94% training dataset and 44.89% for the testing database. Details are presented in Table V below.

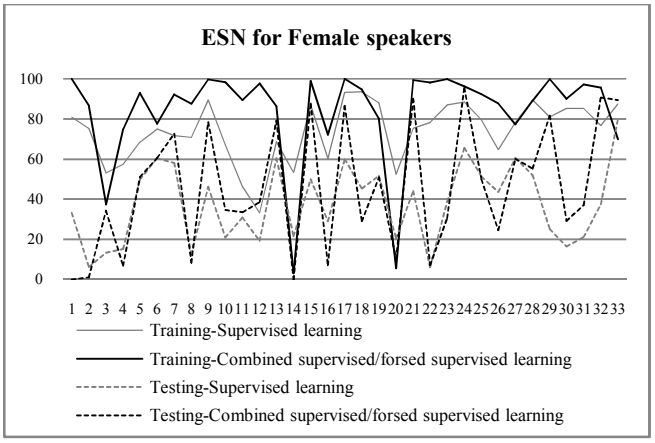

Fig. 6 Average results of phonemes for female speakers' dataset when it trained and tested on the ESN with supervised learning and combined supervised/forced supervised learning algorithms

2. Closed Set Speaker Independent ESN for CSLU Male Speakers

17 males also were randomly selected, and used to train and test on one ESN with the supervised and with one ESN with the combined supervised/ forced supervised learning algorithms. The average results of phonemes for this dataset are presented in Fig. 7.

As you can see from Fig. 7, the ESN with forced learning algorithm can classify most of the phonemes from the CSLU2002 male speakers with an average performance of 83.92% for the training database, and 44.89% with the testing datasets. For the ESN trained with the supervised learning algorithm, the average results of Arabic phonemes recognition was 72.03% training dataset, and 35.43% for testing dataset. Details are shown in Table V.

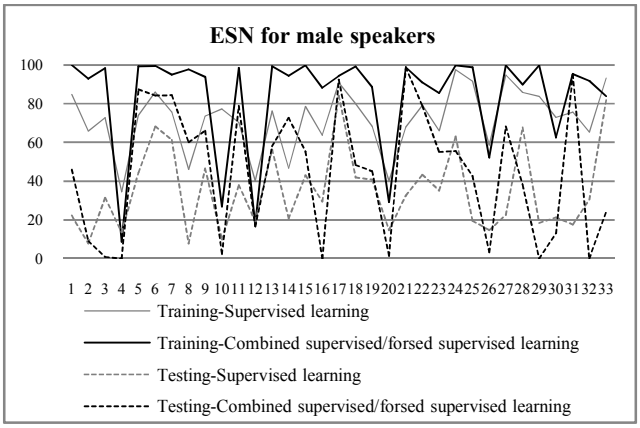

Fig. 7 Average phonemes recognition results for male speakers' dataset when trained and tested on the ESN with supervised learning and combined supervised/forced supervised learning algorithms

3. Closed Set Speaker Independent ESN for CSLU Female and Male Speakers

A dataset of 30 speakers (15 males and 15 females) was constructed from the CSLU2002 females and males datasets and used to evaluate the ESN with supervised and combined supervised/forced supervised learning algorithms. The results of this database are shown in Fig. 8.

It is obvious from Fig. 8 that, using supervised learning, the average Arabic phoneme recognition rate is 51.28% for the females & males training database. In comparison for the same training database using the combined supervised/forced supervised learning algorithm, the average Arabic phoneme recognition rate is 60.50%. For the testing dataset, most of Arabic phonemes in the females & males testing database were recognized with accuracy between 30% and 60% using supervised learning. In contrast, the average recognition of the Arabic phonemes with combined forced supervised/supervised learning was between 50% and 80%.

A comparison between the average results of the CSLU2003 databases using supervised learning and combined supervised/forced supervised learning is shown in Table V.

It is clear from average results of the CSLU2002 female and male speakers' database that, the ESN with combined forced supervised/supervised learning algorithm is best for phonemes classification problems.

The low performance of phoneme recognition in the testing datasets is due to several reasons. The most critical issue is that the speakers are from different countries so have completely different dialectics. Moreover, the limited examples of some phonemes lead to inefficient training for the ESN. Finally, many of the speaker samples contain background noise from the environment; so affecting the ESN performance.

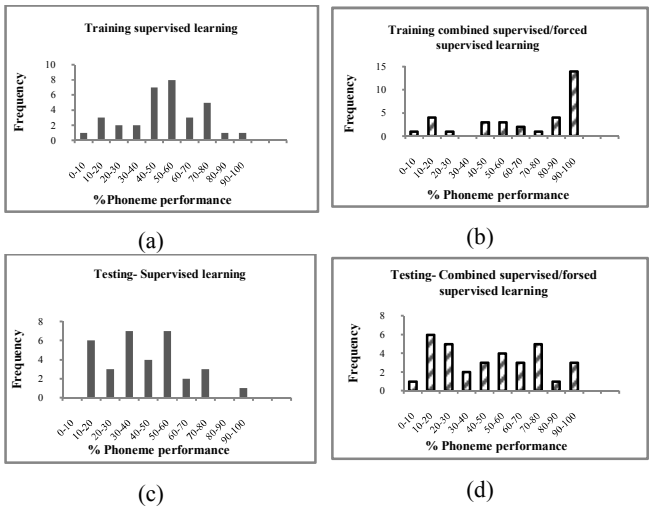

Fig. 8 Histograms for Arabic phonemes recognition performances for the females & males CSLU2002 database, training dataset for supervised learning (a) and combined supervised learning (b), and for testing dataset for supervised learning (c) and combined supervised learning (d)

TABLE V
ESN WITH DIFFERENT LEARNING ALGORITHMS FOR 17 FEMALE SPEAKERS, 17 MALE SPEAKERS, AND A DATASET OF 30 FEMALE AND MALE SPEAKERS

Speakers	Supervised learning		Combined supervised/forced learning	
	Train	Test	Train	Test
Female	73.94	37.72	83.84	45.78
Male	72.03	35.43	83.92	44.89
Female & Male	51.28	33.50	60.50	38.20

V. CONCLUSION

This paper investigates Echo State Networks (ESN) for Arabic phoneme recognition. A novel supervised/forced supervised learning algorithm is proposed that shows improved performance on the CSLU2002 dataset.

In general, speech recognition is a challenge for most standard languages, and Arabic dialectics are no exception. ESN has good performance for the single dialect KAPD dataset. However, a significantly lower performance is obtained for the same ESN when trained with the noisy multi-dialect CSLU2002 dataset. It is suggested that a new Arabic phonetic database for speech recognition purpose should be constructed that addresses the limitations of both the KAPD and the CSLU Arabic databases.

REFERENCES

[1] T. J. Reynolds, C. A. Antoniou, "Experiments in speech recognition using a modular MLP architecture for acoustic modeling, "*Information Sciences*, vol.156, Mar. 2003, pp. 39-54.

[2] W. Chen. S. Chen, C.Lin, "A speech recognition method based on the sequential multi-layer perceptrons, "*Neural Networks*, vol. 9, Nov. 1996, pp. 655-669.

[3] N. Hmad, T. Allen, "Biologically inspired Continuous Arabic Speech Recognition,".*In Research and Development in intelligent systems XXIX,* 32nd ed. Bramer, Petridis Ed. Cambridge, UK: Springer,2012, pp. 245-258.

[4] T. Koizumi, M. Mori, S. Taniguchi, M. Maruya, "Recurrent Neural Networks for Phoneme Recognition,"

[5] M. D. Skowronski, J. G. Harris, "Automatic speech recognition using a predictive echo state network classifier," *Science direct, Neural Networks*, vol. 20, 2007,pp. 414-423.
[6] M. D. Skowronski, J. G. Harris, "Minimum mean squared error time series classification using an echo state network prediction model," *IEEE International Symposium on Circits Systems,* Island of Kos, Greece, 2006, pp. 3153-3156.
[7] M. C. Ozturk, J. C. Principe, "An associative memory readout for ESNs with applications to dynamical pattern recognition, "*Science direct, Neural Networks*, vol. 20, 2007. pp. 377–390.
[8] G. Holzmann, Echo State Networks with Filter Neurons and a Delay&Sum Readout with Applications in Audio Signal Processing., Thesis, Graz University of Technology, Austria, June 2008.
[9] H. Jaeger, H. Haas, "Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless telecommunication," *Science,* vol. 304, 2004, pp. 78-80.
[10] H., Jeager, Adaptive Nonlinear System Identification with Echo State Networks, 2003.
[11] D. Verstraeten, B. Schrauwen, M. D'Haene, D. Stroobandt, "An experimental unification of reservoir computing methods, "*Science direct, Neural Networks*, vol. 20, 2007. pp. 391–403.
[12] M. H. Tong, A. D. Bickett, E. M. Christiansen, G. W. Cottrell, "Learning grammatical structure with Echo State Networks," *Science direct, Neural Networks*, vol. 20, 2007. pp. 424–432.
[13] V. Sakenas, Distortion Invariant Feature Extraction with Echo State Networks, Jacobs University Bremen, Germany, Oct. 2010.
[14] B. Schrauwen, L. Busing, A Hierarchy of Recurrent Networks for Speech Recognition, 2010.
[15] H. Jaeger, M. Lukosevicius, D. Popovici, U. Siewert, "Optimization and Applications of Echo State Networks with Leaky Integrator Neurons," *Science direct, Neural Networks*, vol. 20, 2007. pp. 335–352.
[16] T. P. Schmidt, M. A. Wiering, A. C. van Rossum, R. A.J. van Elburg, T. C. Andringa, B. Valkenier, Robust Real-Time Vowel Classification with an Echo State Network.,2010.
[17] H.J aeger, A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach, International University Bremen, 2005.
[18] I. Sutskever, Training Recurrent Neural Networks, University of Toronto, 2013.