A hybrid neural network based speech recognition system for pervasive environments

A Hybrid Neural Network Based Speech Recognition System for Pervasive Environments

Muhammad Shoaib B. Sehgal, Senior Member IEEE, Iqbal Gondal, Member IEEE, and Laurence Dooley
GSCIT, Monash University, Australia
Shoaib.Sehgal@infotech.monash.edu.au,Iqbal.Gonda@infotech.monash.edu.au, Laurence.Dooley@infotech.monash.edu.au

Abstract

One of the major drawbacks to using speech as the input to any pervasive environment is the requirement to balance accuracy with the high processing overheads involved. This paper presents an Arabic speech recognition system (called UbiqRec), which address this issue by providing a natural and intuitive way of communicating within ubiquitous environments, while balancing processing time, memory and recognition accuracy. A hybrid approach has been used which incorporates spectrographic information, singular value decomposition, concurrent self-organizing maps (CSOM) and pitch contours for Arabic phoneme recognition. The approach employs separate self- organizing maps (SOM) for each Arabic phoneme joined in parallel to form a CSOM. The performance results confirm that with suitable pre-processing of data, including extraction of distinct power spectral densities (PSD) and singular value decomposition, the training time for CSOM was reduced by 89%. The empirical results also proved that overall recognition accuracy did not fall below 91%.

1. Introduction

Speech is the primary and most convenient means of communication between humans [1]. Current human computer interfaces (HCI), like keyboards or mouse are inadequate for ubiquitous/wearable environments. For such environments, speech based inputs are gaining interest because it permits both the hands and eyes to be kept free and therefore less restricted in its use and can achieve quicker communications [2].

The motivation behind this work is to develop a pervasive bioinformatics environment where a speech engine is used as the human-machine interface. For this particular study, Arabic language data [3] has been used to test the efficiency of the speech recognition engine. To develop a continuous Arabic speech recognition system, the input speech is segmented into phonemes using suitable segmentation techniques [4] such as Blind Speech Segmentation [5], Energy Based End Point Detection [6], Zero Crossing Rate and techniques based on Phonetic and Acoustic cues [7, 8]. The novel phoneme-based recognition engine that is presented in this paper is then used for classification. This paper will focus particularly on the classification of Arabic phonemes [9], which is an especially challenging task due to the highly glottal and contextual dependency of the language.

In the proposed classification system Self Organizing Maps (SOM) are used as classifiers. SOM are characterized by a vector space comprising different patterns that exist in the input data space. These vector spaces are developed based on the excitory and inhibitory behavior of the output neurons in SOM [10]. A single neuron or a group of neurons in the output layer contributes to a distinct input in time and space that results in classification and statistical data extraction. This feature has been exploited in this paper to facilitate accurate Arabic phoneme identification.

SOM have a wide range of applicability to complex real world problems ranging from speech recognition to optical character recognition [11]. Kohonen [12] discussed visualization of machine states such as transformers through the application of SOM. Kohonen also identified several important application domains: such as texture analysis and classification, robotics, telecommunication, designing, measuring and testing methods for SOM.

Several studies have shown that CSOM perform better than simple SOM due to their weight optimization for a specific class. Neagoe and Ropot

0-7803-8680-9/04/$20.00 ©2004 IEEE.

[13] applied CSOM for face recognition and multispectral satellite imagery and reported that CSOM had a far greater recognition rate compared to a single SOM, simple Neural Networks and Bayes classifiers. Díaz, et al [14] developed a recognition system for spoken English decimal digits (from 1 to 9). This system used Perceptual Linear Prediction (PLP) coefficients for constructing the CSOM architecture, which provided an overall accuracy of 66.1%. Samouelian [15] worked on knowledge based approach for English consonant recognition and achieved an accuracy of 73%.

Hidden Markov Models (HMM) have been extensively exploited for speech recognition systems. Somervuo [16] used SOM to identify 350 Finnish words, through incorporating competitive HMM-based learning, with a best recognition rate of 90%. Wooters and Stolcke [17] investigated the use of Multiple Pronunciation Models (MPM) for Speaker Independent speech recognition (SISR) systems. The automatic data-driven MPM construction was accomplished by using structural HMM Induction Algorithm. The resulting MPMs were jointly trained with a multi-layer perceptron functioning as a phonetic likelihood estimator. An average recognition accuracy of 74% was reported. Yuk and Flanagan [18] developed a hybrid system based on Neural Networks and HMM for telephonic speech recognition which achieved an overall accuracy of 62%, though in all cases, the disadvantages of HMM are its computational intensity and long training sequences. Smart devices usually possess far lower processing power and memory capability compared with PCs, so the use of HMM in pervasive environment is not feasible [19]. Normally when using HMM in speech recognition systems, each HMM is trained on a phoneme and these phonemes are assembled to form the starting HMMs for words. The methodology used involves monitoring both memory and computational requirements. Separate SOM have been developed and validated for each phoneme which is subsequently assembled for words with segmentation algorithms [4, 5, 6, 7, 8]. It is for this reason that a SOM has a much lower computation and memory requirements than HMM and our empirical results showed that with SVD, a computational time saving of 89% is achieved.

Although a number of phoneme identification studies have been carried out for many modern languages, no research has been reported in context of ubiquitous systems, which use Arabic language as the input tool. This may be due to the fact that Arabic is only the 6th most widely used language, so more emphasis is given to more commonly used languages like English and Mandarin. Also, it is very hard to develop Arabic speech recognition system as compared to English due to the fact that pronunciation is dependent on context and it has Bi-joins, Tri-joins and some times N-joins in between the words. The problem addressed in this paper is to accurately recognize Arabic phonemes. Two approaches are proposed; firstly we can use a single SOM with the same number of output neurons as the number of phonemes to be recognized. The weight optimization involved however will be very complex because when one SOM is used for the classification of multiple classes, a global spread and layer dimensions must be selected, which should be generic for all classes, which may not result in the optimized weight vectors for the specific class. The second approach is to develop a CSOM in which each SOM is responsible for identifying a particular phoneme. A detailed analysis of this approach is provided in this paper.

The rest of the paper is organized as follows: The development of novel hybrid multi-layered Arabic phoneme identification is presented in the paper in Section 2. The system is based on Power Spectral Densities (PSD), singular values, self-organizing maps and pitch contours of the sound waves [3]. The basic principles of SOM are also presented in this Section along with the importance of pitch contours in the recognition system. Section 3 explains the hybrid recognition algorithms developed, with the results obtained from the experiments conducted using the hybrid algorithm given in Section 4. A discussion and some conclusions are presented in Section 5.

2. Hybrid Speech Recognition System

To recognize the consonants, the PSD [20] of the input speech signals are computed with maximum frequency of 8 kHz. In general, consonants are very difficult to identify in the time-domain because of the variation in noise levels and speaker dependent properties in the speech signal. In order to extract the dominant frequencies using PSD, several time windowing approaches were evaluated including Hanning, Hamming, Bartlett, Welch and Gaussian [21]. Hamming and Hanning windows perform better for tonal languages like Mandarin and English, but for Arabic phonemes, performance is superior for the Gaussian window, due to the fact that Arabic is a glottal language and has fewer high frequency components compared with tonal languages.

To obtain the PSD using a Gaussian window, the sampled speech signal S is split into overlapping segments (windows) each with the Gaussian window vector. The coefficients of the Gaussian window are

calculated using equation (1). The length of the window is N, k is the sample index and G is the output signal.

$$G(k+1) \quad e^{x} \tag{1}$$

Where $x = -\frac{1}{2}\left[a(\frac{k-N/2}{N/2})\right]^{2}$, $0 \le k \le N$ and $\alpha \le 2$

A frequency resolution of 20Hz is used for the PSD with zero-padding, so it is an accurate estimate of the short-term, time-localized frequency content of S. In the PSD the time increases from left to right and frequency from bottom to up (ranging from dc to 8 kHz). The average length of S is 61,000 samples and the PSD is a complex matrix with average size of 4000 x 16.

The singular values SV are calculated from the PSD matrix, which is m x n matrix and decomposed into three matrices given by:

$$X = UST^{*} \tag{2}$$

such that $UU^{T} = VV^{T} = I$. Here U and V are two unitary matrices and S is a diagonal matrix containing singular values of X in descending order. Since every matrix has a unique set of singular values therefore this uniqueness is exploited in developing different recognition systems. The advantage is the reduction in computational time and memory requirements as demonstrated later in the paper.

Concurrent Self Organizing-Maps (CSOM) was trained on the first ten SV values. SOM algorithm [10] is based on the principle of winner takes all, which keeps certain biological similarity with the cortical maps. The input vector for SOM is SV (first 10 singular values), and weights between the input layer and the maps are w, the winning neuron k is:

$$k = \min_{i,j \in \text{dimension of } p} \left\{ \left\| SV - w_{ij} \right\| \right\} \tag{3}$$

This particular neuron excites the neurons in its neighbourhood according to the Mexican hat function given by:-

$$C(k_i, k_j, t) = \exp\left(\frac{\left\| k_i \quad k_j \right\|^{2}}{(\alpha(t) S_a)^{2}} \right) \tag{4}$$

where S_a is the number of neurons per dimension, k_i is the winner neuron, k_j is the neighbour of winning neuron and $\alpha(t)$ is the learning rate.

Hebb's learning algorithm for SOM is now applied. This postulates that a synaptic connection is more efficient when the pre-synaptic firing and the post-synaptic firing occur simultaneously as shown in (5) and (6).

For the winner neuron domain

$$\frac{\partial w_{ij}}{\partial} \quad \alpha(t)(SV_{ij} - w_{ij}) \tag{5}$$

For other neurons

$$\frac{\partial w_{ij}}{\partial} = 0 \tag{6}$$

Some researchers have categorically stated that using pitch [22] as a recognition parameter is not a good choice, due to its speaker dependency in developing Speaker Independent Speech Recognition (SISR) systems [23]. Significant research however has also shown that pitch can be used to increase the accuracy of recognition systems. Kitaoka et al [24] worked on glottal sound source features and concluded that glottal features like pitch can be used for SISR systems. Wong and Chang [25] worked on the effects of pitch and lexical tone on different Mandarin speech recognition tasks and found that by considering the tone contexts and incorporating pitch feature lead to higher recognition accuracy. Similarly Chen and Chang [26] developed a recognition system based on Dynamic HMM (DHMM) using pitch values. The results showed that the DHMM achieves approximately a 10% relative error reduction both in base-syllable and tonal syllable recognition tasks. The research presented in this paper also supports this fact and uses pitch as a post-processing layer within the hybrid structure proposed for recognition. The results discussed in Section 3 and 4 show a significant increase of 19% in the overall recognition accuracy.

3. Implementation of Speech Recognition Algorithm

The mathematical model detailed in Section 2, was simulated using MATLAB 6.5.1. The complete phoneme recognition algorithm is defined in Figure1. The input phoneme is processed and the PSD calculated in steps 1 and 2. The singular values are then extracted from the PSD (step 3) and used by the CSOM architecture. The recognition system iteratively computes the Euclidean distance E between the input vector and all SOMs present in CSOM. If the distance is less than the empirical threshold ξ, then this particular phoneme i is a *candidate phoneme*, and all such candidate phonemes are added to vector PID (steps 4 to 9). If no candidate phoneme is identified then the system is unable to recognize the sound, otherwise the phoneme is identified based on the similarity of the standard SOM response and the response of the input signal. If a unique identification of the input sound wave is not obtained, then the

conflict is resolved by activating a pitch analyzer, shown in steps 13 through 16.

The recognition system was trained for 28 basic Arabic phonemes [9] on 100 sound samples for each phoneme. The input sound data was obtained from [20], with 70% of the recorded sounds used for training the SOM. Two layered pre-processing was performed before the training of SOM. In the first layer, PSD values were calculated to facilitate the recognition of consonants from the input speech. As mentioned in Section 2, for spectrogram calculations, the maximum frequency = 8 kHz, frequency resolution = 20 Hz and a Gaussian windowing function was used in (1). The second pre-processing layer implements the SVD to capture the prominent features of respective PSD values (Step 3 in Figure 1). The SVD analysis shows that the first 10 singular values effectively represent the PSD, so m = 10 in Step 3 of Figure 1. Several SOM were developed for each phoneme in the development phase and were tested for accuracy against different phonetic sounds. The final SOM for each phoneme was selected based on the individual performance in terms of percentage recognition accuracy.

```
1. S <- Get utterance
2. PSD <-Calculate PSD from S
3. SV <- Apply SVD on PSD and take 1: m singular
values
4. For i=1: N
5.        E[i] <- Euclidean distance of SOM i
6.        If E[i] < ξ
7.                  Make E[i] part of CE
8.                  Add i to PID
9.        End
10 If size of CE ==0
11.       Phoneme cannot be recognized
12. Else
13.       For j = 1: size of CE
14.                 Sim[j] <-Compare the similarity
                    between pitch contour P[j] of
                    phoneme PID[j] and sound S
15.       End
16. PhonemeId <- PID of max (Sim)
```

Figure 1: Recognition algorithm

4. Experimental Results

Once the individual SOM were optimized for highest possible classification rate, all SOM were integrated to form a concurrent architecture. This arrangement of CSOM was extensively tested for validation of the data set. It was noted that certain phonemes were misclassified, resulting in an overall decrease of recognition accuracy from 91.7% to only 71.9%. The results for CSOM are shown in A2 in Table 1. The recognition accuracy of /a:/ which was previously recorded as 100% (A1 in Table1) dropped to only 56.25% (A2 in Table 1). This reduction in accuracy was due to the misclassification of /a:/ as /H/ (CP in Table 1). Similar discrepancies were identified for the /b/, /l/ and /z/ phonemes.

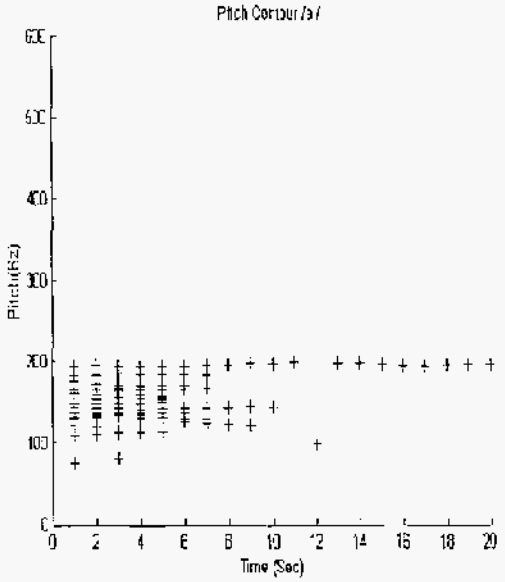

Figure 2: Pitch of /a:/ for 18 people

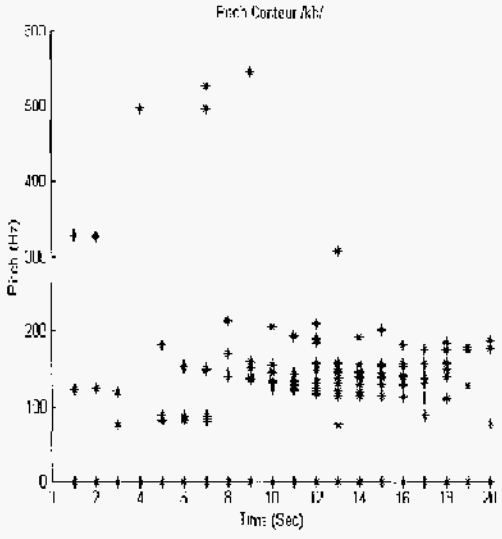

Figure 3: Pitch of /kh/ for 18 people

In order to solve this misclassification problem, a post processing layer was added, which compared the standard pitch contours with pitch contours of the input sounds. Experiments showed that distinct pitch contours were present for most of the misclassified phonemes when tested for all the speakers. For example, in the case of /kh/, there was no pitch in the initial 5 frames, whereas a continuous band of pitch was observed for /a:/ during the same time period. This is shown in Figures 2 and 3. Similarly /d/ has continuous pitch in the initial frames while /H/ has no pitch in this region. Phoneme / / has continuous pitch in the initial frames as opposed to /H/. Therefore, any misclassification between /a:/ - /H/, /d/ - /X/ and / / - /H/ can be resolved using the pitch information. Similar analysis was conducted for all the phonemes and their misclassifications. A pitch analyzer compared the standard pitch contours and the pitch contours of the input sound was added as a post-processing layer in the hybrid system, resulting in an overall recognition accuracy up to 90.8% as shown in A3 of Table 1.

The training and recognition times of the SVD-based recognition system were recorded and compared with the non-SVD based recognition system i.e., the SOM were directly trained on the PSD. The experiments confirmed an improvement in the CPU throughput from 80.35% to 89.48% in both training and recognition.

5. Conclusions

This paper presents a hybrid Arabic phoneme recognition system for pervasive environments, based on PSD, singular values, self-organizing maps and pitch contours. The study indicates that training and recognition time of CSOM has been dramatically reduced due to the introduction of SVD. With the introduction of pitch contours as a post processor, recognition accuracy increased from 71% to above 90%, confirming the judgment to use the pitch features in phoneme recognition for various phonetic sounds. An overall recognition accuracy of 90.84% was observed with reduction in training and recognition time by a factor of 80.38% and 89.48% respectively. This recognition accuracy compares very favorably with the performance of other systems such as those identified in [3, 13, 14, 15, 17].

6. References

[1] B.H. Juang, and S. Furui, "Automatic Recognition and Understanding of Spoken Language – A First Step Towards Natural Human-Machine Communication", *IEEE*, 2000, pp. 1142-1165.

[2] S. Furui, K. Iwano, S. Hori, T. Shinozaki, Y. Saito, and S. Tamura," Ubiquitous Speech Processing", *Proceedings IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP'01)*, Salt Lake City, 2001, vol.1, pp.13-16.

[3] M. Shoaib, M. Awais, S. Masud, S. Shamail, and J. Akhtar, "Application of Concurrent Generalized Regression Neural Networks for Arabic Speech Recognition", *The 2nd IASTED International Conference on Neural Networks and Computational Intelligence*, Grindelwald, Switzerland, 2004.

[4] R. Martinez, A. Alvarez, P. Gomez, M. Perez, V. Nieto, and V. Rodellar, "A Speech Pre-processing Technique for End-Point Detection for Highly Non- Stationary Environments", *Eurospeech'97*, 1997.

[5] M. Sharma and R. Mammone, "Blind Speech Segmentation: Automatic Segmentation of Speech without Linguistic Knowledge", *The Fourth International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, 1996.

[6] K. Bush, A. Ganapathiraju, P. Kornman, J. Trimble, and L. Webster, "A Comparison of Energy- Based Endpoint Detectors for Speech Signal Processing", *IEEE SouthEastCon'96*, 1996.

[7] A. Weber, "The Role of Phonetics in the Segmentation of Native and Non-Native Continuous Speech", *Workshop on Spoken Access Processes*, pp. 143-146.

[8] A. Weber, "Phonotactic and Acoustic Cues for Word Segmentation in English", *International Conference on Spoken Language Processing (ICSLP'00)*, 2000.

[9] "Novel Speech Recognition Models for Arabic", Johns-Hopkins University Summer Research Workshop, pp. 7-8, 2002, 876—880,
http://www.halcyon.com/pub/journals/21ps03-vidmar.

[10] T. Kohonen, "Physiological Interpretation of the Self-Organizing Map Algorithm", *Neural Networks*, 1993, vol. 6, pp. 895- 905.

[11] H.H. Song, and S.W. Lee, "A Self Organizing Neural Tree for Large Pattern Classification", *IEEE Third International Conference on Document Analysis and Recognition (ICDAR'95)*, 1995.

[12] T. Kohonen, "New Developments and Applications of Self- Organizing Maps", *IEEE International Workshop on Neural Networks for Identification, Control, Robotics, and Signal/Image Processing (NICROSP'96)*, 1996.

[13] V.E. Neagoe and A.D. Ropot, "Concurrent Self-Organizing Maps for Pattern Classification", *First IEEE International Conference on Cognitive Informatics (ICCI'02)*, 2002.

[14] F. Díaz, J. M. Fernández, P. Gómez, V. Rodellar and V. Nieto, "Spoken-Digit Recognition using Self-organizing Maps with Perceptual Pre-processing", *International Work Conference on Artificial and Natural Neural Networks*, 1997.

[15] A. Samouelian, "Knowledge Based Approach to Consonant Recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, 1994.

[16] P. Somervuo, "Self-Organizing Maps for Signal and Symbol Sequences", *PhD Thesis Helsinki University of Technology, Neural Networks Research Centre*, 2000.

[17] C. Wooters and A. Stolcke, "Multiple-Pronunciation Lexical Modeling in A Speaker Independent Speech Understanding System", *International Conference on Spoken Language Processing (ICSLP'94)*, 1994.

[18] D. Yuk and J. Flanagan, "Telephone Speech Recognition using Neural Networks and Hidden Markov Models", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)*, 1999.

[19] D. Devisch, "Building Speech Recognition in Portable Products", *Multimedia and DSP, Speech Recognition, Electronics Engineer*, 1999.

[20] V.W. Zue and L.F. Lamel, "An Expert Spectrogram Reader: A Knowledge-Based Approach to Speech Recognition", *IEEE Acoustic Speech, Signal Processing*,
Tokyo, Japan, pp. 1197-1200, 1986.

[21] K. Vinay and J.G. Proakis, *Digital Signal Processing using MATLAB*, BookWare Companion Series, 2000.

[22] P. Boersma, "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound", *Institute of Phonetic Sciences (IFA), Proc.17*, pp. 97-110, 1993.

[23] Jerry Liu, "Effects of Pitch Tracking Features in Mandarin Speech Recognition with Wide Range of Accents", http://www.ee.columbia.edu/~jliu/e6820/dialect/pitch_project.pdf.

[24] N. Kitaoka, D. Yamada, S. Nakagawa "Speaker Independent speech recognition using features based on glottal sound source", *International Conference Spoken Language Processing (ICSLP'02)*, pp. 2125-2128, 2002.

[25] Wong and Chang, "The Effect of Pitch and Lexical Tone on Different Mandarin Speech Recognition Tasks", *Eurospeech'01*, Scandinavia, 2001.

[26] Chen and Chang, "A New Dynamic HMM Model for Speech Recognition", *Eurospeech'01*, Scandinavia, 2001.

[1]**Layer Dim:** SOM Layer Dimensions.
[2]**CP:** Conflicting Phonemes when SOM were tested for all phonemes.

Appendix

Table 1
Recognition accuracies,
A1: %Accuracy of Experiment1 (E1) when SOM were trained to achieve maximum accuracy for individual phoneme, A2: %Accuracy of Experiment2 (E2) when SOM were tested for all phonemes, A3: %Accuracy of Experiment3 (E3) when post processing based on pitch contours was applied

Phoneme		Layer	E1	E2		E3
		Dim[1]	%A1	CP[2]	%A2	%A3
/a:/	ا	5 8	100	/θ/,/sw/.	56.25	100
/b/	ب	5 8	100	/t/	62.5	100
/t/	ت	5 8	100	/ /, /n/	62.5	81.25
/θ/	ث	10 10	81.25	/f/, / /	50	68.75
/ /	ج	5 5	100	/h/,/zw/	56 25	93.75
/H/	ح	5 4	100	/zw/, /n/	56.25	100
/X/	خ	3 5	100	/d/,/zw/	37.5	100
/d/	د	2 8	87.5	/x/,/zw/	81.25	100
/ /	ذ	2 2	100	No conflict	100	100
/r/	ر	4 4	93.75	/tw/	50	100
/z/	ز	16 16	81.25	/l/, /k/	31 25	68.75
/s/	س	3 3	100	/tw/	100	100
/ /	ش	2 9	100	/t/,/r/, /n/	81.25	81.25
/sw/	ص	2 15	81.25	/dw/, /zw/	50	100
/dw/	ض	2 11	81.25	/d/,/zw/	50	50
/tw/	ط	5 1	100	No conflict	100	100
/zw/	ظ	1 20	100	/zw/	93.75	93.75
/ /	ع	12 16	81.25	/d/	81.25	81.25
/ /	غ	7 7	93.75	/θ/	93.75	100
/f/	ف	3 20	87.5	/a:/, /0/	87.5	93.75
/k/	ق	10 16	87.5	/ /	62 5	100
/q/	ك	10 12	100	/s/	68.75	68.75
/l/	ل	15 20	87.5	/ /, /s/	87.5	93.75
/m/	م	12 16	87.5	/q/	87.5	100
/n/	ن	4 8	75	/ /	75	100
/w/	و	8 12	93.75	/q/	93.75	100
/h/	ه	8 10	81.25	/ /, /i:/	81.25	81.25
/ /	ى	9 9	87.5	/b/, /q/	75	87.5
Overall Accuracies			**91.74**		**71.87**	**90.84**