J. King Saud Univ., Vol. 17, *Comp. & Info. Sci*., pp. 1-21 (A.H. 1425/2004)

The Relative Distance Vector Neural Network (RDVNN) Model: A Hybrid Approach to Speech Recognition

Elgasim Elamin Elnima
King Saud University,
College of Computer and Information Sciences,
P.O. Box 51178, Riyadh, Saudi Arabia

(Received 29 June 2003; accepted for publication 11February, 2004)

Abstract. This paper introduces a novel insight to the problem of Automatic Speech Recognition (ASR). Worldwide many practical systems had been developed for ASR. Most of these systems were based on Hidden Markov Models (HMM). This is state-of-the-art paradigm in ASR. Despite the fact that HMMs are successful under a diversity of conditions, they do suffer from some limitations that limit their applicability to real-world noisy environments. As a result, several researchers moved to Artificial Neural Networks (ANNs) as an alternative technique for ASR, in order to overcome the limitations encountered in pure HMM implementation. Soon after, interest moved over to hybrid systems that combine HMMs and ANNs within a single unifying hybrid architecture. In this study a hybrid DTW/ANN ASR system will be introduced, explained, implemented and analyzed, which has been given the name Relative Distance Vector Neural Network (RDVNN) Model.

Adequate experiments had been performed to reveal the main characteristics of the present novel hybrid ASR system. The results are believed to be encouraging and the system is easy to implement. For speaker dependent the accuracy is near perfect (error rate is less than 1%). For speaker independent models the results attained are comparable with most world-wide results known for the state-of-the-art ASR small task systems. Many aspects of the RDVNN technique are illustrated through experimental work to demonstrate these findings. One of the main advantages of the RDVNN method is that it can be applied to various other similar problem domains.

1. Introduction

The early models applied traditional pattern matching techniques to speech recognition. However, faced with the temporal variation in speech signals, most of the early models resorted to dynamic time warping techniques [1-4]. The performance of dynamic time warping techniques proved to be reasonable in the case of speaker independent isolated word systems but their performance in case of the more complex speaker independent continuous recognition systems have been less than satisfactory.

During the 1980's, researchers switched from traditional simple models to more complex statistical models like Hidden Markov Models (HMM's). Despite their improvement over other models, HMM's suffer from a number of limitations and problems.

Towards the late 1980's and being faced by these limitations, some researchers turned to Artificial Neural Networks (ANN's) [4-10]. Work in this area has been mainly motivated by the success of neural networks in the area of pattern recognition in general. It has been hoped that ANN's will help greatly in the basic speech classification process if properly trained.

One of the first problems met by researchers in this direction was the temporal dynamic nature of speech patterns which makes them very different from classical static patterns. To solve this problem a number of models have been suggested. Two important examples of these are the Time Delay Neural Network (TDNN) technique [11] and the recurrent neural network [12, 13].

Although neural networks are good classifiers that can easily generalize and despite the introduction of the TDNN and related models, ANN's did not succeed in providing a general framework for ASR that will take into consideration long sequences of acoustic features. This led to the appearance of hybrid ANN/HMM and ANN/DP ASR models during the early 1990's [14-16]. A number of researchers started to believe that combining ANN's with HMM's or DP will yield the best of these approaches. Examples of these are the models developed by Bourland and Wellekens [12], Bengio [8], Niles and Silverman [14], Tebelskis [7], and Terntin [15]. A survey of these models may be found in [6, 8].

In most ANN/DP approaches, the DP algorithm has been used as a postprocessor [10] to integrate ANN results with some prior knowledge of the temporal structure of the input sequences. In this study a new hybrid ANN/DTW model is introduced, tested and analyzed [18]. The main difference between the present model and the previous models is that the DTW algorithm is used as a preprocessor rather than a postprocessor. Further, the new model relies on a second level feature space based on the traditional first level speech features. The model consists of a DTW front-end and a feedforward ANN backend. The DTW front-end is used to compute a set of relative distance feature vectors representing the time-warped distances between the input utterance (words) and the elements of a chosen reference set (speaker's reference template set). The tests performed proved that the model is robust and accurate.

Test data had been selected from five test data corpora, including TIMIT, CTIMIT [19], and TIDIGITS [20].

The RDVNN model is introduced in the next section. After that the corpora are described followed by the presentation of the experiments and discussions. Finally come the conclusions and the recommendations.

2. Architecture and Operation of the RDVNN:

As shown in Fig. 1 the RDVNN consists of two major components: a front-end dynamic time warping (DTW) component and a backend feedforward neural component. The DTW front-end computes (a second level set of features) a relative distance vector (RDV) based on the input utterance feature vectors. The RDV vectors represent the relative distances between the input utterance feature vectors and the feature vectors of the reference template set. The RDV vectors are then used to train the feeedforward network. The training algorithm used is the backprobagation algorithm.

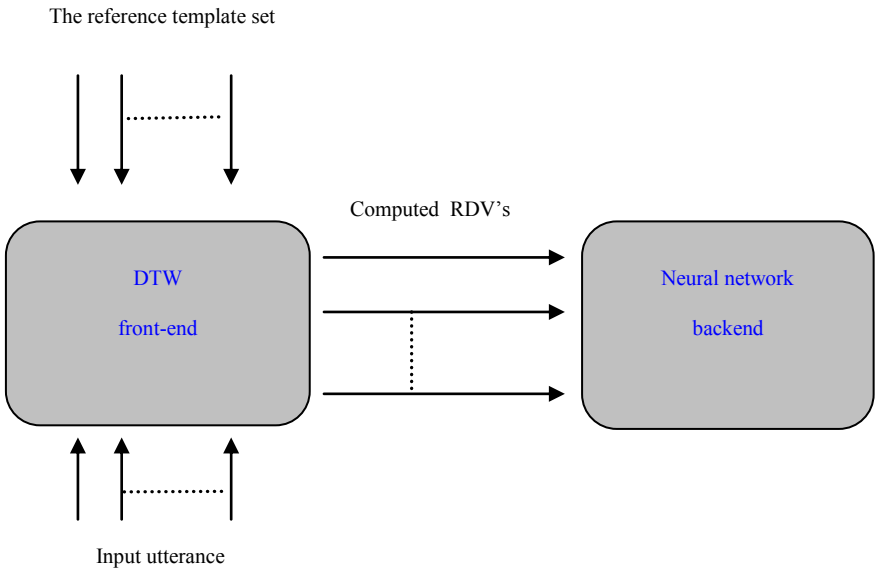

Fig. 1. The relative distance vector neural network (RDVNN) model diagram.

2.1 Computation of the RDV vector

Generally, given a reference template set consisting of N templates (each template representing an utterance) and the input utterance x, the front-end of the RDVNN computes the relative distance between x and each of the reference templates d_k, say, where k is the index of the template (1 <= k <= N). The computed relative distances form a new feature vector of dimension N. The algorithm may be summarized as follows:

```
For  k := 1 to  N  do
   d[k]  :=  DTW(x, Template[k]);
```

where *DTW(x, Template[k])* is a call for the dynamic time warping function.
The RDV vector, computed above, has the same size as the reference template set.

2.2 Training the feedforward component

The back-end neural component, as shown in Fig. 2, is a multi-layer feedforward neural network [6, 10]. The structure and organization of the training data set greatly depends on the purpose for which the recognition system is used. In the case of speaker dependent models and speaker verification systems, the data usually consists of multiple recordings for the utterances pronounced by the same speaker. In this case, one version of the recordings is used as a reference template set, while the remaining recordings are used to compute distance vectors.

As for speaker independent systems, the training set usually consists of multiple recordings for multiple speakers of different ages and gender. In this case, one recording for any of the speakers can be assumed as the reference template set.
The training procedure may be summarized as follows:

Given reference template set consisting of M recordings for all speakers, each recording consisting of N utterances, plus a reference template set for one of the speakers:

1. For each utterance x in the training set use the DTW front-end to compute M RDV vectors.
2. Present the M RDV's to the feedforward neural network component together with the correct output.

2.3 Test and recall

The testing and recall procedure is straightforward and may be summarized as follows:
Given a test utterance u:

1. Use the DTW front-end of the RDVNN to compute the corresponding RDV vector.

2. Present the RDV vector to the feedforward component which should return the expected utterance.

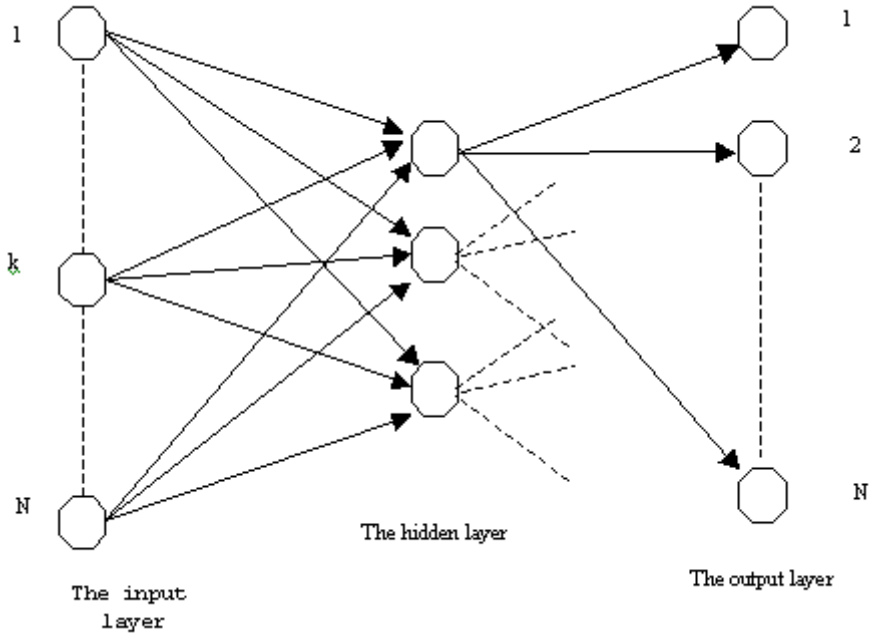

Fig.2. Back-end feedforward neural component of the RDVNN.

3 Speech Databases

Five sets of speech databases have been used to investigate the performance of the RDVNN. All five corpuses have been used for speaker independent tests, but the KSULAD corpus has been used for speaker dependent tests as well.

3.1 Local arabic digits corpus (KSULAD)

The Local Arabic Digits corpus consists of 1050 utterances recorded by five male speakers. Each speaker has been asked to record 21 repeats of each of the ten digits from

0 to 9, pronounced in Arabic. The general properties of the KSULAD corpus are summarized in Table 1 below:

Table 1. Local arabic digits (KSULAD) corpus characteristics

General characteristics	
Speech corpus	Locally recorded Arabic digits
Speakers	5 Males
Utterances	1050 (210 per speaker)
Speech file format	Windows .wav format
Sampling rate	11025 Hz
Sample resolution	8 – bit

To train and test the RDVNN, the wave form utterances would be transformed into Mel-frequency Cepstral Coefficients (MFCC) feature vectors whose parameters are shown in Table 2:

Table 2. Local arabic digits (KSULAD corpus) MFCC specifications

MFCC specifications (KSULAD corpus)	
DC component	DC component is estimated over all the utterance and subtracted.
Sampling frequency	11025 Hz
Reemphasis	None
Window size	256 sample
Spectral analysis	FFT
Frequency warping	Mel-scale
Mel-fliter banks	32
Mel-filter shape	Triangular
Cepstral features	12 + (energy sometimes)
Cepstral filtering	None
Feature normalization	In limited number of experiments

3.2 The Otago university corpus

The Otago university corpus consists of 630 utterances recorded by 21 speakers. Each speaker has been asked to record 3 repeats of each of the 10 digits from 0 to 9, pronounced in English. The general characteristics of speech data in the Otago corpus are given in table 3 below:

Table 3. The Otago university digits corpus characteristics

Otago Data Set (Digits Corpus) , New Zealand http://kel.otago.ac.nz/hyspeech/corpusinfo.html	
General characteristics	
Speech corpus	Otago English Digits Corpus
Speakers	Total =21; 11 Male , 10 Female
Utterances	630 (30 per speaker)
Speech file format	Windows PCM(raw) format
Sampling rate	22050 Hz
Sample resolution	16 – bit signed
Training set	
Speakers	15 speakers [1:4, 7, 9, 11, 12, 16:21]
Test set	
Speakers	4 Speakers [5, 10, 13, 15]

The MFCC feature parameters for the Otago university corpus are similar to the ones given in Table 2 except that the window size is 512 samples.

3.3 The TIMIT corpus

The TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. All utterances of sentences SA1 and SA2 are used in this study. The complete training set consists of 462 speakers, and the complete test set comprises 168 speakers.

The Sampling Frequency is 16000 Hz, other parameters are similar to the ones given in Table 2.

3.4 The cellular TIMIT speech corpus (CTIMIT)

It is important to evaluate the performance of RDVNN technique with data recorded in noisy and non-ideal environments like CTIMIT corpus [19]. CTIMIT (Version 1.0 alpha, February, 1996) was developed by Lockheed-Martin Sanders, Inc. The CTIMIT corpus is a cellular-bandwidth adjunct to the TIMIT Acoustic-Phonetic Continuous Speech Corpus (NIST Speech Disc CD1- .1/NTIS PB91-505065, October 1990). CTIMIT has 8 kHz sampling rate, but other parameters are similar to the ones given in Table 2.

3.5 The TIDIGITS corpus

The data used is a single-word TIDIGITS (ISIP) speech that consists of 880 training utterances, and 880 test utterances, down sampled to 8 kHz, with 40 men and 40 women. MFCC parameters are similar to the ones given in Table 2.

4. The RDVNN in a Detailed Example

In this section, the RDVNN technique is explained in more depth making use of a detailed example from the TIMIT corpus.

We will start at the sentence wave file (sa1.wav) spoken by the second speaker in the training set of the TIMIT corpus, his abbreviated name is "ADC0" and he is a male from dialect area 3. The wave file has a number of related files. One of them is sa1.wrd, which lists the word labeling of the sentence. In the above mentioned sentence the word GREASY is the 7th word, which has a length of 6953 samples. Transforming the wave form of the word GREASY to MFCC parameters (12 parameters) generates 26 frames, each frame contains 12 MFCC parameters. The number of elements in the 26 frames is 12 x 26 = 312 elements, compared to 6953 samples for the wave form.

The last stage of preprocessing is the computation of DTW distances to produce the RDV vector. This process would produce a vector of length equal to the number of words to be recognized. Now each word will be represented by an RDV vector of 11 parameters only. The artificial neural network will be trained on this optimized RDV vectors (see Fig. 3).

Figure 3 graphs the RDV vectors for the word (Water) in sentence sa1 for the first seven speakers (ADC0, ADD0, AEB0, AEM0, AEO0, AFM0, and AFM0) from the TIMIT corpus. The discriminative nature of the RDV vectors is clear from this figure.

5. Experimental Framework

In the following sections we would present and discuss a series of experiments that demonstrate the robustness of the RDVNN technique. The experiments are based on two types of models. The first type trains all words to be recognized in one neural network model (one model for all words). The second type uses independent model for each word (per word model).

The network topology generally consists of three layers, the first layer is the input layer and its number of nodes is, usually, equal to the dimension of the RDV vector. The second (hidden) layer has a number of nodes that can be varied through a series of values for testing purposes. The output layer is kept constant at a number of nodes that equals two for the per-word model, or equal to the number of utterances to be recognized when the one model for all words is used.

The maximum number of epochs (5 – 40), and goal (0.05 – 0.001), are set according to the needs of the experiment under test.

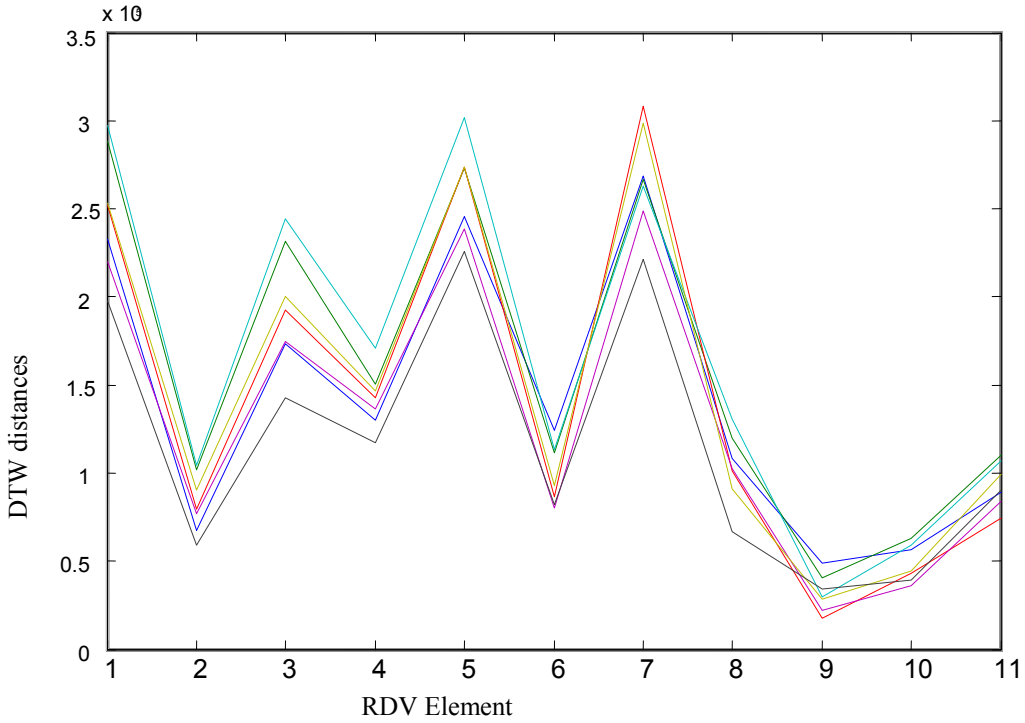

Fig. 3. The word WATER: 7 speakers (TIMIT corpus).

Unless otherwise stated, first speaker data set was used as reference template set. Also, a reference template set might be selected according to one of the optimization procedures discussed in sections 6.4 and 6.5. In all neural network training, one of the following configurations (training methods) had been used.

5.1 Network training methods

Three methods for training the neural network have been adopted, these are briefly described below:

(i) The all-data training method

In this method all the training sets would be used for training in one huge batch.

(ii) Small batches training method

In this method, the training data is divided into small batches. Training takes place by giving each batch a number of epochs (e.g. 2 epochs). This process is then repeated for all batches in the set. The length of the batch can be varied. A suitable batch size can be determined through experimentation. This method is used for large corpuses, e.g. the TIMIT.

(iii) Multistage training method

In this method, the network is trained in two stages or more. The first stages are similar to the batch training method, and the last stage is an all-data training method. This combination proves to be useful in reducing the training time.

The goal of the first stage may be different from the goal of the second stage (i.e. 0.005 for the first stage and 0.001 for the second Stage)

6. Results of Experiments and Discussions

6.1 KSULAD and Otago corpuses

A short account will be given for the local and Otago data sets. The local data set (KSULAD) achieved nearly 100% recognition rate for speaker dependent case and achieved an average of 98.5% recognition rate for speaker independent case.

Otago data set achieved over 95% average recognition rate for speaker independent case, which can be improved to around 99% recognition rate when reference template set optimization is used (sections 6.4 and 6.5). The above results are based on a one model for all words network configuration.

6.2 TIMIT corpus (one model per word)

This experiment used a model for each word (11 models) training method. The DTW applied to training and test data produced 73% recognition rate. The summary of 18 runs is shown in Table 4.

6.3 Improved template experiments

In this process we select a common template that is contributed by one or more speakers. The aim of this is to find a better template which can produce more recognition rate at DTW stage and consequently may produce higher recognition rate at the final stage. The search is limited to a subset of speakers; we select candidates every 16th speaker restricting the search to 29 speakers from the total training set of 462 speakers. A network model is trained for each of the 11 words separately. The same test is repeated for different hidden layer values. The output layer is kept at 2 nodes. Two criteria for selecting such an improved template have been investigated, these are:

1. Simple counting method, and

2. Discriminative method.

Table 4. Recognition rate summary (18 runs) (the output layer is 2 nodes)

Net topology, hidden layer	Train%	Test%	Average (test)
7	99.363	99.188	
7	99.766	99.449	**99.392**
7	99.837	99.538	
8	99.814	99.606	
8	99.828	99.567	**99.570**
8	99.810	99.538	
9	99.857	99.597	
9	99.857	99.523	**99.572**
9	99.887	99.597	
10	99.834	99.577	
10	99.837	99.597	**99.571**
10	99.843	99.538	
11	99.875	99.587	
11	99.889	99.562	**99.565**
11	99.830	99.547	
12	99.864	99.577	
12	99.871	99.606	**99.590**
12	99.866	99.587	
Average	**99.818**	**99.543**	**99.543**

6.3.1 The simple counting template selection method

In this process, for each word, we select a speaker who attains the highest recognition rate in the DTW stage when compared to other speakers in the set. The template produced by this method was given the name "bestc". The average DTW recognition for this template was 81.15%, compared to 99.622% average recognition for the network part. Table 5 presents the network simulation summary for five runs.

6.3.2 Discriminative template selection method

We design this process in the hope that the reference template set selected has some discriminative characteristics. i.e. templates that give low DTW distances when compared against words of the same class and give high distances when compared with words of other classes. This feature is approximated by the ratio:

qRatio = Sum1 / sum2 ;

Where:

sum1: Sum of the squares of the DTW distances for the words of the same class,
sum2: Sum of the DTW distances for the words that do not belong to the tested class.

Table 5. Summary for template selection methods bestc & bestq

Hidden layer size	simple counting method (bestc)		discriminative method (bestq)	
	Train%	Test%	Train%	Test%
9	99.835	99.597	99.889	99.646
10	99.844	99.606	99.955	99.715
11	99.894	99.636	99.902	99.685
13	99.909	99.631	99.936	99.656
14	99.916	99.641	99.966	99.695
Average	**99.880**	**99.622**	**99.930**	**99.679**
Max	**99.952**	**99.641**	**99.966**	**99.715**
Min	**99.835**	**99.597**	**99.889**	**99.646**

In this process, for each word, we select the template of the speaker who attains the lowest qRatio when compared to other speakers in the set. However, by implementing this technique we can increase the discrimination of the DTW stage. This approach proved to give better results than the former simple counting approach. The reference template sets produced by this method were given the name “bestq”. The DTW recognition for these templates was 84.9%, compared to 99.679% average recognition for the network part (see the last two columns on Table 5).

6.4 Using two reference template sets

In this experiment the RDV vector was based on two speakers' reference template set, which will double the size of the RDV vector.

A network model is trained for each of the 11 words separately. The same test is repeated for a series of hidden layer values. The output layer is kept at 2 nodes.
The results depicted in Table 6, show that this approach achieves a small improvement.

In conclusion the two template approach improves the average recognition by about 0.2%, but it has the drawback of increasing processing time and also doubles the input vector size.

Table 6. Summary for the two reference template sets method

Net topology, hidden layer	Train %	Test %
7	99.916	99.739
10	99.930	99.744
12	99.973	99.779
14	99.966	99.798
17	99.968	99.779
18	99.971	99.788
19	99.957	99.774
20	99.968	99.784
21	99.980	99.813
22	99.971	99.793
Average	**99.960**	**99.779**

6.5 The first 12 speakers' templates of the TIMIT corpus

In this experiment we explore the effect of choosing a reference template arbitrarily. The template set was selected for each one of the first 12 speakers of TIMIT corpus in turn. Here, in this experiment, all speakers attain DTW recognition between 52% and 74% with an average of 64.86%. On the other hand the network recognition rate is always greater than 99.2% for all speakers, with a maximum of 99.48%, and an average of 99.39%.

From these results it is obvious that:

- Higher DTW recognition rate did not always lead to a higher network recognition gain.
- The DTW recognition rate varies extremely depending on speaker's reference template set used. On the other hand the corresponding network results are rather stable.

6.6 One network model for all words

Network models can be designed separately for each word. Alternatively, all words can be trained and recognized by one network model (classification). In this experiment, the one model for all words we run 11 times with the number of nodes of the hidden layer varied across a number of settings. The network model is trained for all the 11 words in one common model. The template set used is bestq which was described in section 6.3.2. For the Network topology, the hidden layer had been tested over node values (11 , 12 , 13, 15, 17, 22), while the output layer is fixed at 11 nodes. The results were depicted in Table 7.

Table 7. Summary; TIMIT train(462) test(168); template (bestq)

Network hidden layer size	Train%	Test%
11	99.075	98.431
11	98. 839	98.647
11	99.134	98.593
12	99.174	98.377
12	99.193	98.000
13	98.800	98.052
13	99.351	98.864
15	99.154	98.323
17	99.271	98.268
22	99.859	98.106
22	98.406	98.214
Average	**99.149**	**98.311**
Max	**99.859**	**98.864**
Min	**98.406**	**98.000**

These results show that high recognition was achieved by this type of models which means that we can use any of the two models and the choice between them would depend on other factors like stability and execution time.

6.7 Tests on TIMIT with a vocabulary of 21 words

Previous experiments use a limited vocabulary set of 11 words. In this experiment the vocabulary size is extended to 21 words instead, including both sentences SA1 and SA2, and will incorporate all data from the training set (462 speakers) and testing set (168 speakers) of the TIMIT corpus. A network model is trained for each of the 21 words separately. The output layer is kept at 2 nodes. The results are summarized in Table 8.

Table 8. Summary for 21 words of sentences SA1 and SA2 of the TIMIT

Net topology, hidden layer	Train %	Test %
4	99.536	99.333
5	99.615	99.393
10	99.767	99.424
11	99.774	99.402
17	99.823	99.449
Average	**99.703**	**99.4002**

These results reassure the high recognition rate even when the vocabulary size is doubled.

6.8 Cellular TIMIT speech corpus (CTIMIT)

Tests with CTIMIT corpus will let us know the extent to which the RDVNN technique works with cellular collected databases. This experiment uses 21 words including both sentence SA1 and sentence SA2 of the CTIMIT corpus. The training data contains 143 speakers, and the test set data contains 44 speakers.

A network model is trained for each of the 21 words separately. The same test is repeated 10 times for each of the first five speakers in the training set of the CTIMIT corpus. For the Network topology, the hidden layer was fixed at 8 nodes, while the output layer is kept at 2 nodes. The results are summarized in Table 9.

Table 9. Summary of DTW and RDVNN recognition Rates for the first 5 reference speakers, Training speakers = 143 , Test speakers = 44

No	Reference speaker	DTW recognition rate		DRVNN recognition rate	
		Train %	Test %	Train %	Test %
1	ADD0	25.708	26.407	99.326	97.820
2	AEB0	16.517	17.641	99.351	97.722
3	AEM0	27.306	29.437	99.286	97.939
4	AEO0	33.899	33.983	99.477	97.939
5	AFM0	40.892	43.182	99.534	98.320
Average		**28.864**	**30.130**	**99.395**	**97.948**
Max		**40.892**	**43.182**	**99.534**	**98.320**
Min		**16.517**	**17.641**	**99.286**	**97.722**

The minimum recognition rate is 97.9%. Examination of the depicted results shows that the RDVNN technique performs well in these types of environment. Certainly there is some degradation in recognition rate from the corresponding TIMIT data.

6.9 TIDIGITS corpus experiment

In this experiment, tests were carried on the 10 English digits. A one model for all words was used throughout these tests. Table 10 presents recognition results where the training and testing data is for the same gender. Table 11 presents recognition results for men and women.

Table 10. TIDIGIT same gender recognition

Hidden layer size	Train MEN Test MEN		Train WOMEN Test WOMEN	
	Train%	Test%	Train%	Test%
20	99.500	98.750	99.750	98.500
20	99.250	98.750	99.750	99.000
25	99.000	98.750	100.000	98.750
25	99.250	99.500	99.500	98.250
30	99.250	100.000	99.500	98.750
30	99.750	99.500	100.000	98.500

Average	**99.333**	**99.208**	**99.750**	**98.625**

Table 11. TIDIGIT recognition for Men and Women, first Speaker's Templates first & second Speakers' Templates

Hidden layer size	**First speaker's templates**		**First & second speakers' templates**	
	Train%	Test%	Train%	Test%
20	97.875	96.625	99.250	98.500
20	97.125	96.500	99.125	98.375
25	97.375	96.500	99.875	98.625
25	97.625	96.750	99.250	98.500
30	97.750	97.250	99.625	98.875
30	97.250	96.750	99.750	98.750
Average	**97.500**	**96.729**	**99.479**	**98.604**

Table 10 shows that the RDVNN can give 99.208% average recognition rate for training and testing with men, and a corresponding 98.625% average recognition for women with women. Speaker independent testes on men and women (Table 11) attain 96.729% average recognition rate for one speaker reference template set, and 98.604% for two speaker reference template set.

Average recognition time per word is around 40 us. Training time can be as low as two minutes for a data of 400 utterances, and might reach up to one hour for a data of more than 5000 utterances.

7. Conclusions and Recommendations

7.1 Conclusions

This paper presents a novel low complexity hybrid recognition system for isolated words in clean or noisy environments. The proposed technique implements a DTW front-end followed by a neural network backend. Several experiments have been performed to demonstrate the effectiveness of the present RDVNN technique. The findings of these experiments are as follows:

i) The best test data performance of the RDVNN technique was achieved with the selection of an optimum template of one more speaker's template set when training with a limited number of speakers. However, large corpuses like the TIMIT the improvement is less that 0.2% (section 6.3 and 6.4).

ii) Speaker dependent configuration can attain a near perfect recognition rate of 100% in most of the runs in contrast to an average recognition of 85% for the DTW method as revealed by experiments on the KSULAD corpus.

iii) It can be deduced from the results of " TIMIT corpus a one model per word" experiment (section 6.2) that speaker independent configuration when tested on

11 words of TIMIT can attain average recognition rate of 99.606% compared to average recognition of 73% for the DTW method.

iv) Otago data set achieved over 95% average recognition rate for speaker independent case, which can be improved to around 99% recognition rate by the selection of an optimum template set.

v) Fine tuning of the network training parameters like batch size produce good training results within a shorter time.

vi) Usually experiments on the TIMIT corpus are run with one template. Results of these runs showed that any arbitrarily selected template can produce acceptable recognition rate with only very small discrepancies between the different results (Using Two Reference Template sets experiment, section 6.4). When two reference template sets are used, this approach improved the average recognition by about 0.2%, but it has the obvious drawback of increasing processing time and also doubles the input vector size.

vii) Form "The First 12 Speakers' Templates of the TIMIT Corpus" experiment (section 6.5), it is observed that:, higher DTW recognition rate did not always lead to a higher network recognition gain. The DTW recognition rate changes considerably depending on speaker template, but the corresponding network results did not follow these broad changes.

viii) Tests on the two sentences SA1 and SA2 (21 words) of the TIMIT achieved an average recognition of 99.4%. In contrast, the average for DTW recognitions is 64.67% (section 6.7).

ix) The common model for all words results (depicted in Table 7) shows that high recognition was achieved by this type of model which means that we can use any of the two types of models and the choice between one of them would depend on other factors such as stability and execution time.

x) From the results of "TIDIGITS corpus Experiment" experiment (section 6.8) it is seen that the minimum recognition rate for cellular environment (CTIMIT) was 97.72%, and a maximum of 98.32%, with the average of 97.95%. The corresponding DTW recognition rates are 17.64%, 43.18%, and 30.13% respectively. Also, investigations of the depicted results reveal the fact that the RDVNN technique performs well in the cellular environment. Surely, there is a small degradation in recognition rate from the corresponding TIMIT data.

xi) In the TIDIGITS experiments, Table 10 shows that the RDVNN can give 99.208% average recognition rate for training and testing with men, and a corresponding 98.625% average recognition for women. Speaker independent tests on men and women (Table 11) attain 96.729% average recognition rate for one speaker reference template set and 98.604% for two speaker reference template sets.

xii) The RDVNN technique is language independent because it is based on matching of relative distances between patterns.

7.2 Recommendations

Results obtained in this study clearly show that the RDVNN technique, as developed in this experiment is very efficient in small vocabulary of isolated words. However, the technique is very flexible and can easily be extended to other problems. This study has several aspects that require profound investigation both in terms of different application domains and modification of the existing techniques. Some of these are listed below:

i) Normalization of the input vector may have some effect on the recognition accuracy.
ii) DTW implantations come in a variety of forms. All other forms can be examined.
iii) Acoustic features used are only the basic ones. Sophisticated acoustic feature extraction methods might be considered.
iv) RDVNN technique may be applied to many other problem domains where DTW is used.
v) Other methods that are a substitute for DTW (e.g. Linear Time Alignment and Trace Segmentation (TS)) can play a role within the RDVNN technique. Though, the attained accuracy may be a little different.

References

[1] Myers, C. S. and Rabiner L. R. "Connected Digit Recognition Using A Level-building DTW Algorithm." *IEEE Trans. on Acoustics Speech and Signal Processing. ASSP-29,* No.3 (June 1981), 351-363.

[2] Rabiner, L. R. and Schmidt, C. E. "Application of Dynamic Time Warping to Connected Digit Recognition." *IEEE Trans. on Acoustics, Speech and Signal Processing*. *ASSP*-28, No.4 (August 1980), *377-388* .

[3] Rabiner, L. R.., Rosenberg, A. E. and Levinson, S. E. "Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition." *IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-26*, No. 6 (December 1978), 575-582.

[4] Sakoe, H. and Chiba, S. "Dynamic Programming Algorithm Optimization For Spoken Word Recognition." *IEEE Trans. on Acoustics, Speech and Signal Processing. 26, No. 1, February* (1978), *43-49.*

[5] Lippmann, Richard P. "An Introduction to Computing with Neural Nets." *IEEE ASSP Magazine*, April (1987).

[6] Morgan, D. P., and Scofield C. L. *Neural Networks and Speech Processing.* Kluwer international series in engineering and computer science, 1991.

[7] Tebelskis, J. *Speech Recognition Using Neural Networks.* PhD. Dissertation, School of Computer Science, Carnegie Mellon University, 1995.

[8] Bengio, Y. *Neural Networks for Speech and Sequence Recognition*. London: International Thompson Computer Press, UK, 1996.

[9] Fausett, L. *Fundamentals of Neural Networks. Englewood Cliffs.* NJ: Prentice Hall, 1994.

[10] Creaney, M. J. and Gorgui-Naguib, R. N. "A Scaly Artificial Neural Network Architecture for Speaker Independent Isolated Word Recognition Using Non-Linear Time Alignment." *Proc. of The IEEE Int. Conf. on Neural Networks, ICNN '94.* VII, Orlando (1994), 4431-4436.

[11] Waibel, A. "Modular Construction of Time-delay Neural Networks for Speech Recognition." *Neural Computation*, 1 (1989), 39-46.

[12] Bourlard, H. and Wellekens, C. J. "Speech Dynamics and Recurrent Neural Networks." *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. ICASSP '89, Glasgow (1989), 33-36.

[13] Lang, K. J. and Hinton, G. E. *The Development of the Time-delay Neural Network Architecture for Speech Recognition*. Technical Report CMU-CS-88-152, Carnegie-Mellon University, 1988.

[14] Niles, L. T. and Silverman, H. F. "Combining Hidden Markov Model and Neural Network Classifiers." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1 (1990), 417-420.

[15] Trentin, E. *Robust Combination of Neural Networks and Hidden Markov Models for Speech Recognition.* PhD Thesis, Universita' di Firenze (DSI), Feb 27, 2001

[16] Huang, W. Y. and Lippmann, R. P. "HMM Speech Recognition with Neural Net Discrimination." In: Lippmann, R. P., Moody, J. E. and Touretzky, D. S. (Eds), *Advances in Neural Information Processing Systems 3*, San Mateo, CA: Morgan Kaufmann, (1991), 194-202.

[17] Bourlard, H. and Wellekens, C. "Links Between Hidden Markov Models and Multiplayer Perceptrons." *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 12 (1990), 1167-1178.

[18] Elnima, E. E. *The Relative Distance Neural Network: A Hybrid Model for Automatic Speech Recognition*. PhD Thesis, University of Khartoum, 2003.

[19] Brown, K. L. and George, E. B. "CTIMIT: A Speech Corpus for the Cellularenvironment with Applications to Automatic Speech Recognition." In: *Proc. IEEE Int'l Conf. on Acoust., Speech and Signal Processing*, (May 1995), 105-109.

[20] ISIP. The Institute for Signal and Information Processing at Mississippi, State University. http://www.isip.msstate.edu/projects/speech

نموذج الشبكات الذكية للبيانات النسبية : طريقة هجينة لتمييز الكلام

القاسم الأمين النعمة

قسم تقنية الحاسب، كلية علوم الحاسب والمعلومات، جامعة الملك سعود

ص ب ٥١١٧٨، الرياض ١١٥٤٣، المملكة العربية السعودية

(قدّم للنشر في ٢٠٠٣/٠٦/٢٩م؛ وقبل للنشر في ٢٠٠٤/٠٢/١١م)

ملخص البحث. أصبح استخدام الحاسب الآلي يشمل كثيراً من المجالات العلمية والتطبيقية في مختلف نواحي الحياة. ومن ضمن هذه الاستخدامات استخدام الحاسب لتميز الكلمات والأصوات. ولقد شهد هذا المجال تطوراً مطرداً حتى غدت نسبة التميز في كثير من هذه الأنظمة عالية نسبيا مما مكن دخولها مرحلة التطبيق الفعلي.

أنظمة التميز الآلي للكلمات يمكن تقسيمها إلى أقسام رئيسية، نذكر منها: الأنظمة التي تعتمد على مقارنة النماذج (DTW) والأنظمة التي تستخدم الطرق الإحصائية (HMMs) وقسم ثالث يقوم على استخدام الشبكات الذكية (ANNs) التي تحاكي أسلوب عمل العقل في بعض جوانبه. كما يمكن الربط و التوفيق بين هذه الطرق الرئيسية للخروج بأنظمة هجينة (Hybrid) تجمع كثيراً من الصفات الإيجابية لهذه الأنظمة.

من خلال هذه الدراسة تم تطوير نظام تميز للكلمات المفردة النطق يجمع بين مقارنة النماذج وأنظمة الشبكات الذكية للوصول إلى نظام يحقق قدراً مناسباً من الدقة في تميز الكلمات والأصوات.

تم اختبار النظام على عدة قواعد بيانات صوتية منها ما هو محلي (الأرقام العربية) وقاعدة بيانات من جامعة Otago في نيوزلندا (الأرقام الإنجليزية) وقاعدة بيانات TIMIT والتي أخذت منها كلمات منفردة من جملتين أساسيتين و توسعت الاختبارات لتشمل CTIMIT و TIDIGITS .

إن هذا النظام المطور يتمتع بكثير من الميزات الإيجابية؛ نذكر منها: سهولة التطبيق؛ الدقة العالية؛ الاستقرار في وجود عوامل مختلفة؛ الفاعلية في بيئات مختلفة مثل الهاتف الجوال. وإضافة لكل تلك الميزات يمكن تطبيق هذا النظام في مجالات أخرى مشابهة. مثل الأنظمة التي تعتمد على مقارنة النماذج (DTW) .

ولقد أطلق على هذا النظام اسم Relative Distance Vector Neural Network (RDVNN).