“ЖУРНАЛ РАДИОЭЛЕКТРОНИКИ” N 7, 2009 оглавление

SPEAKER INDEPENDENT SPEECH RECOGNITION SYSTEM USING NEURAL NETWORKS

N.Uma Maheswari[1], A.P.Kabilan[2] , R.Venkatesh[3]

[1]Senior Lecturer, Dept. of CSE, P.S.N.A College of Engg& Technology, Dindigul-624622,India
[2]Principal, Chettinad college of Engg& Technology, Karur-639114,India
[3]Senior Lecturer, Dept. of CSE, R.V.S College of Engg& Technology, Dindigul-624005,India

Received July 4, 2009

Speaker independent speech recognition is important for successful development of speech recognizers in most real world applications. While speaker dependent speech recognizers have achieved close to 100% accuracy, the speaker independent speech recognition systems have poor accuracy not exceeding 75%.In this paper we describe a two-module speaker independent speech recognition system for all-British English speech. The first module performs phoneme recognition using two-level neural networks. The second module executes word recognition from the string of phonemes employing Hidden Markov Model. The system was trained by British English speech consisting of 2000 words uttered by 100 speakers. The test samples comprised 1000 words spoken by a different set of 50 speakers. The recognition accuracy is found to be 92% which is well above the previous results.

Keywords: speaker independent speech recognition, Neural Network, Hidden Markov Model, phonemes.

НЕЗАВИСИМАЯ СИСТЕМА РАСПОЗНАВАНИЯ РЕЧИ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Н. У. Махешвари[1], А. П. Кабилан[2], Р. Венкатеш[3]

[1]P.S.N.A Инженерно-технический колледж, Диндигул, Индия
[2]Инженерно-технический колледж, Карур, Индия
[3]P.V.S. Инженерно-технический колледж, Диндигул, Индия

Получена 4 июля 2009 г.

Независимая от говорящего система распозновния речи важна для успешной разработки устройств распознавания речи для большинства реальных приложений. В то время как зависящие от говорящего распознаватели речи достигли точности 100%, системы распознавания, независящие от говорящего, имеют низкую точность, не превышающую 75%. В работе описывется двух-модульная независящая от говорящего система распознавания речи для англоязычной речи. Первый модуль выполняет распознавние фонем, используя двухслойную нейронную сеть. Второй модуль выполняет распознавание слов, исходя из строки фонем, используя скрытую марковскую модель. Система была протестирована англоязычной речью, состоящей из 2000 слов, произнесенных 100 говорящими. Точность распознавания составила 92%, что намного выше результатов предыдущих работ.

Ключевые слова: независящее от говорящего распознавание речи, нейронная сеть, скрытая марковская модель, фонемы.

I. INTRODUCTION

Automatic speech recognition is a process by which a machine identifies speech. The machine takes a human utterance as an input and returns a string of words , phrases or continuous speech in the form of text as output. The conventional methods of speech recognition insist in representing each word by its feature vector and pattern matching with the statistically available vectors using HMM or neural network. In this work we have adopted a bottom-up approach, in which a speech signal is resolved into a string of phonemes and phoneme recognition forms the basis for word recognition which in turn constitutes the full text output. string of words or phrases which in turn constitute the output in the form of text.

2.SYSTEM ARCHITECTURE FOR SPEECH RECOGNIZER

The proposed speech recognition method comprises of three steps: acoustic signal processing, phoneme recognition and word recognition (Figure 1). First, we digitize the input speech waveform phoneme-by-phoneme. Phonemes are recognized using artificial neural network (high level and low level) and subsequently words are recognized from the clusters of phonemes using HMM.

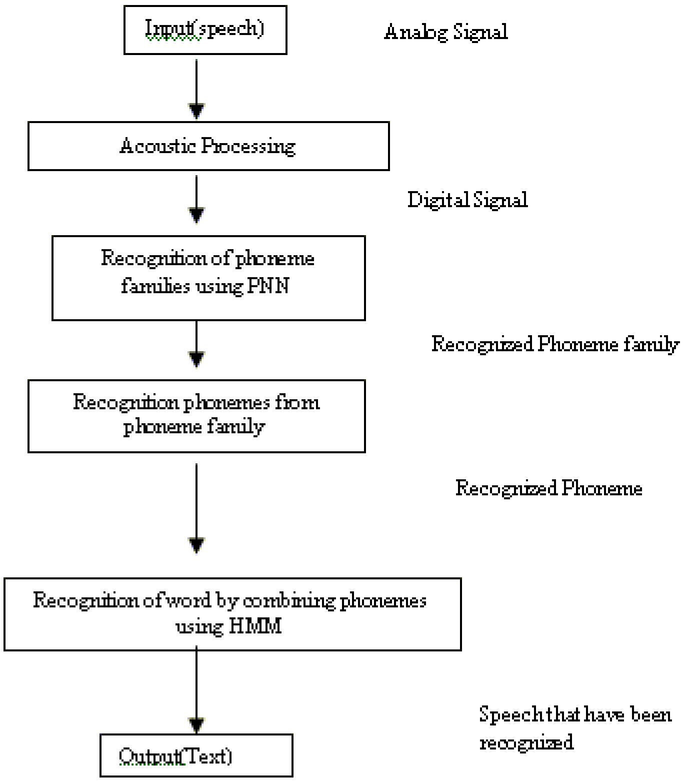

Input(speech)
Analog Signal
Acoustic Processing
Digital Signal
Recognition of phoneme families using PNN
Recognized Phoneme family
Recognition phonemes from phoneme family
Recognized Phoneme
Recognition of word by combining phonemes using HMM
Speech that have been recognized
Output(Text)

Fig.1 System architecture for speech recognizer

3. PHONEME RECOGNITION USING NEURAL NETWORKS

This paper proposes a modular-based classifier for the problem of phoneme recognition which in turn is used for speech recognition. A phoneme is the smallest meaningful distinguishable unit in a language’s phonology. Since the total number of phonemes for each language is finite, the goal of phoneme recognition is to classify a speech signal into phonemes with a given set of speech features. We apply a two level classifier method to design the phoneme recognition system. It uses both statistical and neural network based methods in a hierarchical modular system.

We use the concept of phoneme families. To obtain phoneme families, we employ k-mean clustering method[16]. A given unknown phoneme is first classified into a phoneme family at high level classification using Probabilistic Neural Networks(PNN). Due to the powerful characteristics of probabilistic neural networks such as rapid training, convergence to the Bayesian classifier and generalization, we use this statistical-based classifier to construct an initial topology of the proposed hierarchical modular system[3],[4]. Then, the exact label of the phoneme is determined at low level classification using Recurrent neural network(RNN).Multilayer perceptron(MLP) and recurrent neural network (RNN) are employed as local experts to discriminate time-invariant and time-variant phonemes, respectively. In the second module the words are recognized from time-isolated string of phonemes and the string of recognized words are displayed as text in the output using the Hidden Markov Model

3.1. PROPOSED PHONEME RECOGNITION SYSTEM

We propose a new modular-based approach for phoneme recognition problem. This method consists of two levels of classification: high and low. We define various phoneme families in the high level classification. To obtain the phoneme families, clustering techniques such as *k*-mean clustering are applied. To find an appropriate number of phoneme families, we consider different values of *k*. The value yielding a lower classification error rate is chosen as the best value. Here we have taken the total phoneme families at low level, i.e. *k* = 7.A typical architecture of the proposed system is presented in Fig.2 with k=3

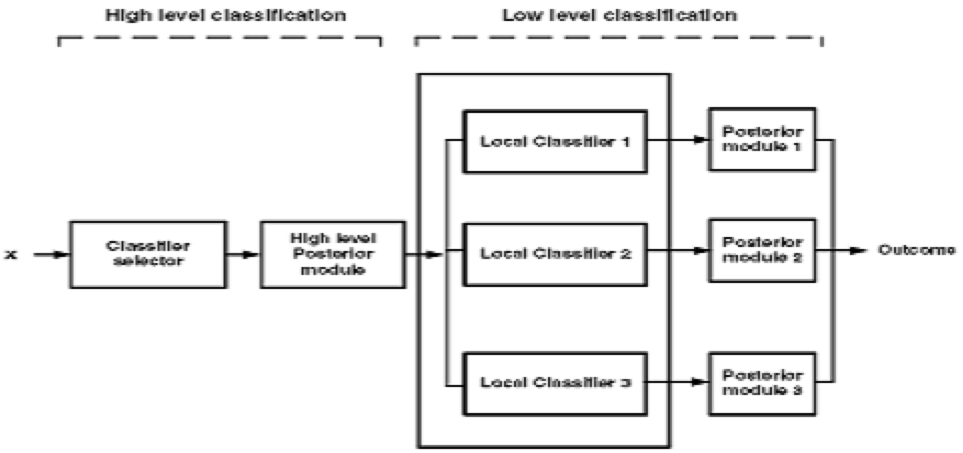


Fig. 2.A typical architecture of the proposed system

There are several components in the system comprising classifier selector, high level posterior module, low level classifiers, and low level posterior modules. An unknown input acoustic waveform is digitized with 38 samples (which was found to be the optimum were fed into the high level classifier. Here we use input layer with 38 nodes, two hidden layers with 12 and 6 nodes each and an output layer with 7 nodes. The high level classifier recognizes an unknown phoneme *X* as phoneme family k^* as follows:

$$k^* = arg \max_k \sum_{l=1}^{\ell} DM_{high}(l,k) \; ; \; k = 1, 2, ..., K; \tag{1}$$

where,

$$DM_{high}(l,k) = \begin{cases} 1 & \text{if } k = arg \max_k p(f_l | CL_k) \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

where, p(fl/CLk) stands for the posterior probability of the *lth* window of frame-level data given phoneme family *CLk*.
Also, l denotes the number of windows of phoneme.
In other words, phoneme data is fed to the high level classification on window-by-window basis. Posterior probability of each window of frame-level data given all phoneme families is obtained through high level classifier. Then, label of the pattern is estimated by aggregating the responses over all windows using high level posterior module.For each phoneme family, we designate a classifier to perform low level classification. Suppose that the output of the high level classification for input pattern *X i*s denoted by *CLk*. Hence, the corresponding low level classifier classifies *X* as member $j^*$of *CLk* if

$$j^* = arg\max_j \sum^{\ell} DM^k_{low}(l,j) \; ; \; j = 1,2,...,m_k; \qquad (3)$$

where,

$$DM^k_{low}(l,j) = \begin{cases} 1 & \text{if } k = arg\max_j p(f_l|CL_{k,j}) \\ 0 & \text{Otherwise} \end{cases} \qquad (4)$$

This approach requires two different algorithms which are learning and classification. The learning algorithm expressed in Algorithm 1, identifies the proper topology of the system. The classification algorithm presented in Algorithm 2,classifies any unknown phoneme.

3.2. PNN AND NEURAL NETWORKS

PNN as statistical classifier is applied to determine the initial topology of the system. PNN is also used to recognize silence at low level.According to maximum a posteriori (MAP) probability, an unknown pattern *X* is classified as class *Ci,* if

$$P(X|C_i)P(C_i) \geq P(X|C_j)P(C_j) \quad j =i \qquad (5)$$

where, *P*(*X/Ci*) denotes a probability distribution function of class *Ci* and *P*(*Ci*) is a prior probability of class *Ci*. MAP provides a method for optimal classification. It clarifies how to classify a new sample with the maximum probability of success given enough prior knowledge. As can be seen from (5), to classify an unknown pattern *X*, probability distribution function (*pdf*) of all classes should be known. Having enough training data, one is able to estimate such a *pdf*.

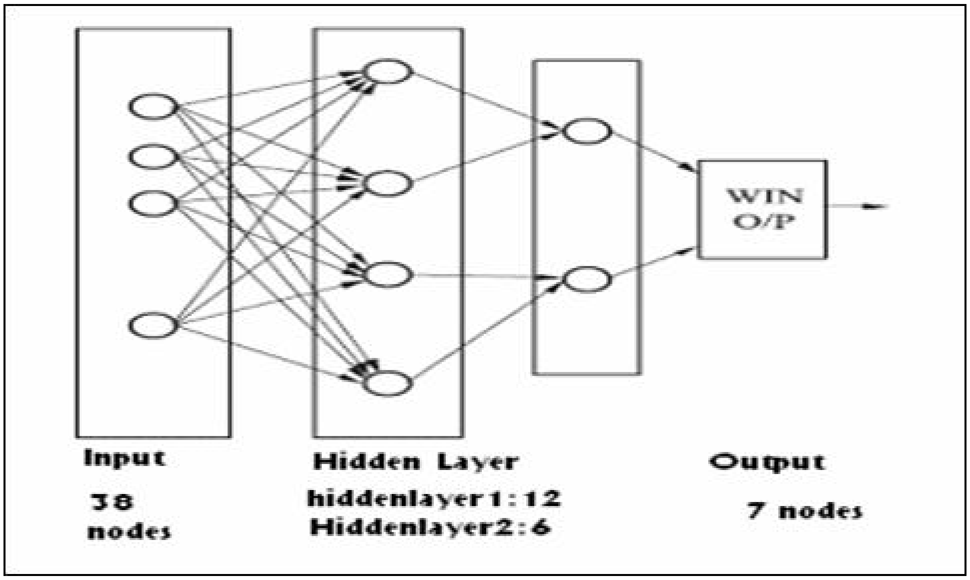


Fig. 3.The architecture of PNN

PNN employs Parzen window density estimation with Gaussian windowing function as a *pdf* estimator given by:

$$P(X|C^c) = \frac{1}{(2\pi)^{\frac{q}{2}}\sigma^q}\frac{1}{n_{C^c}}\left(\sum_{i=1}^{n_{C^c}} exp\left(\frac{-(X-Y^i_{C^c})^T((X-Y^i_{C^c}))}{2\sigma^2}\right)\right) \quad (6)$$

where σ is a smoothing parameter which represents smoothness degree of the probability distribution function and *q* shows the dimension of the feature space. $Y^i_{C^c}$ denotes training data i of class *Cc*. Also, *nCc* denotes the number of training data in class *Cc* and *T* is the vector transpose. The probabilistic neural network provides a four-layer network to map an unknown pattern *X* to any number of classes based on (5)[3]. Exemplar is meant for identification of phonemes and class is for classification of phonemes from exemplar.

3.3. RNN AND NEURAL NETWORKS

MLP and RNNare used as local experts at low level classification in the present modular system[1]. The different phoneme families are considered at low level classification and it identifies the individual phoneme from the identified phoneme families. Each family contains a set of phonemes which are similar in terms of speech features.

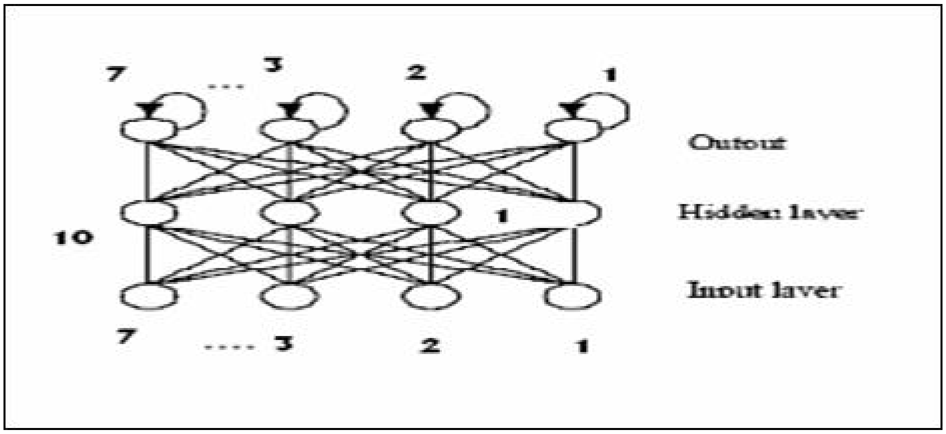


Fig. 4 RNN Architecture

In other words, the phonemes which are very *close* to each other in terms of Euclidean distance are grouped in the same family. Therefore, there is a distributed structure in which each module or family is responsible to learn some specific part of the problem and give its response during the recognition phase. To design such expert modules capable of recognizing any upcoming phoneme pattern, we need more powerful discriminators in low level classification. Multi-layer perceptron is suited for the recognition of phonemes with time invariant input parameters . Also, RNN can learn the temporal relationships of speech data and is capable of modeling time-dependent phonemes . Since both MLP and RNN are trained on other classes' data they are able to discriminate between similar classes. The structure of the used MLP and RNN are very similar, except that RNN has feedbacks on its output layer. The input of the networks is a window of frame level features. Both MLP and RNN have as many outputs as the number of phonemes, *N*.
We can denote the output by

$$O = (o_1, o_2, ..., o_{N-1}, o_N) \qquad (7)$$

where, for a given input *x* belonging to phoneme *k*.
we use this architecture for speech recognition especially for speech recognition by using Backpropagation Through Time (BPTT) as learning algorithm. This architecture also has been proved that this architecture better than MLP in phoneme recognition accuracies by using Backpropagation algorithm. The Backpropagation Through Time (BPTT) algorithm is based on converting the network from a feedback system to purely feedforward system by folding the network over time. The network can then be trained if it is one large feedforward network with the modified weights being treated as shared weight. The weights are updated either after iteration or after the final iteration of the epoch.

4. SYSTEM TOPOLOGY'S PARAMETERS

The initial topology of the proposed system is determined using probabilistic neural network. In this regard, a number of parameters including smoothing parameter (σ), phoneme families number (*k*) and window size of frame-level data (*w*) are clarified.

Algorithm 1 Learning algorithm

```
1: Provide a set of training data for each phoneme
2: Find sample mean of training data belonging to each phoneme C i and denote it as SMC i .
3: for k = 2 to K do
4:Obtain  phoneme families
    $CL^k = \{CL_1^k, CL_2^k, ..., CL_k^k\}$        (8)
  using (k-mean clustering) on  sample    mean data.
```

```
5:Find the best value of smoothing parameter which leads to minimum error rate at
  high level classification. Denote this value $\sigma^{k,1}$.
6:Find the best value of smoothing parameter which leads to minimum error rate at low level classification. Denote this value $\sigma^{k,2}$ .
7:Obtain the overall error rate of the system, $E_k$ , considering $\sigma^{k,1}$.and $\sigma^{k,2}$ as the smoothing parameters of the high and low level classifiers, respectively.
8: end for k9: k ← arg min{E }, where k is the number of the smoothing parameters used.
```

```
Algorithm 2 Classification algorithm
1: Provide a testing pattern $X$ to the network.
2: Compute the output of the high level classification for a testing pattern .
3: Provide testing pattern $X$ to the selected low level classifier.
4: Obtain the output of corresponding low level classifier for testing pattern .
```

We have examined the accuracy of the system in terms of classification rate considering different values for the smoothing parameter at both high and low level classifications. The value which leads to minimum classification error is for small σ near zero, PNN behaves like nearest neighbor classifier.

5. HMM MODEL FOR SPEECH RECOGNITION

The HMM is a probabilistic pattern matching technique in which the observations are considered to be the output of stochastic process and consists of an underlying Markov chain. It has two components: a finite state Markov chain and a finite set of output probability distribution.

Words are usually represented by networks of phonemes. Each path in a word network represents a pronunciation of the word. The same phoneme can have different acoustic distributions of observations if pronounced in different environment.

1. Define a set of L sound classes for modeling, such as phonemes or words; call the sound classes $V=\{v_1,v_2,\ldots\ldots v_l\}$

2. For each class, collect a sizable set (the training set) of labeled utterances that are known to be in the class.

3. Based on each training set, solve the estima tion problem to obtain a "best" model i.i for each class V_i(i = 1, 2, . . . , *L).*

4. During recognition, evaluate P_r (0 / λ_i.,) (i =1, 2, . , L) for the unknown utterance 0 and identify the speech that produced 0 as class

$$\Pr(\mathbf{O} \mid \lambda_j) = \max_{1 \le i \le L} \Pr(\mathbf{O} \mid \lambda_i). \qquad (9)$$

5.1. HMM CONSTRAINTS FOR SPEECH RECOGNITION SYSTEMS

HMM could have different constraints depending on the nature of the problem that is to be modeled. The main constraints needed in the implementation of speech recognizers can be summarized in the following assumptions:

5.1.1 First order Markov chain :

In this assumption the probability of transition to a state depends only on the current state

$$P(q_{t+1}=S_j/q_t=S_i\,,\,q_{t-1}=S_k\,,\,q_{t-2}=S_w\,,\,\ldots\ldots\,,\,q_{t-n}=S_z) \approx P(q_{t+1}=S_j/q_t=S_i) \qquad (10)$$

5.1.2 Stationary states' transition

This assumption testifies that the states transition are time independent, and accordingly we will have:

aij = P(qt+1=Sj / qt=Si) for all t (11)

5.1.3 Observations independence*:*

This assumption presumes that the observations come out within certain state depend only on the underlying Markov chain of the states, without considering the effect of the occurrence of the other observations. Although this assumption is a poor one and deviates from reality but it works fine in modelling speech signal.

This assumption implies that:

$$P(O_t/O_{t-1}, O_{t-2}, \ldots\ldots O_{t-p}, q_t, q_{t-1}, q_{t-2} \ldots\ldots q_{t-p}) = P(O_t/ q_t, q_{t-1}, q_{t-2} \ldots\ldots q_{t-p})$$

(12)

where p represents the considered history of the observation sequence.Then we will have :

bj(Ot) = P(Ot/qt=j) (13)

5.1.4 Left-Right topology construction

aij = 0 for all j > i+2 and j < i

$$\pi_i = P(q_1 = S_i) = \begin{cases} 1 & \text{for} \quad i=1 \\ 0 & \text{for} \quad 1<i\leq N \end{cases}$$

$$(\text{ i.e. } \pi = \{1 \ \ 0 \ \ \ldots\ldots\ldots. \ \ 0 \ \})$$

5.1.5 Probability constraints:

Our problem is dealing with probabilities then we have the following extra constraints:

$$\sum_{j=1}^{N} a_{ij} = 1$$

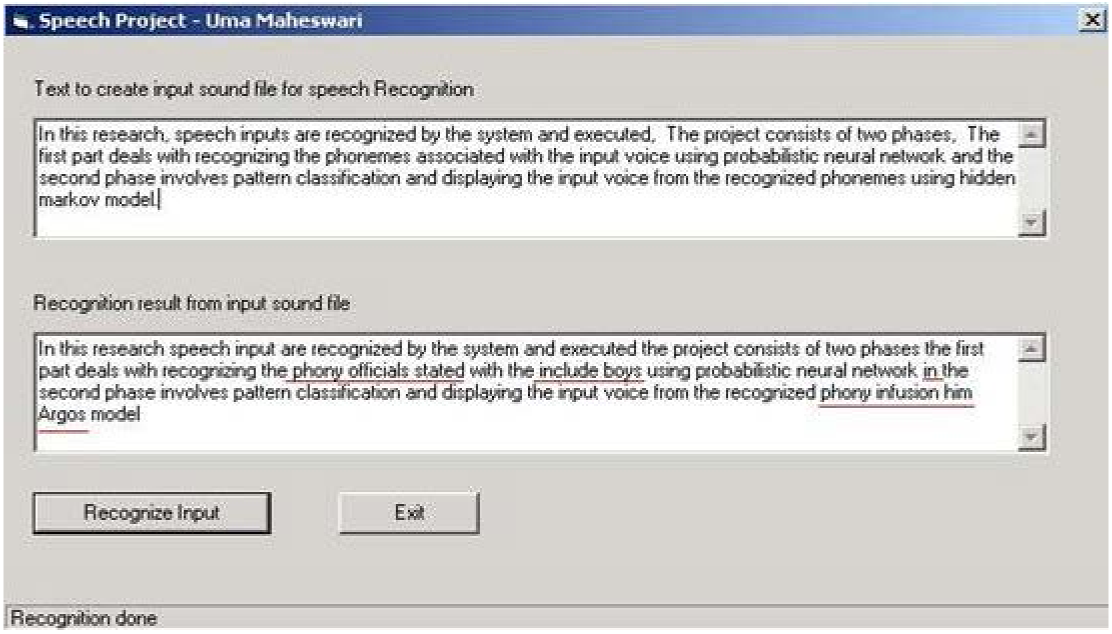
$$\sum_{j=1}^{N} \pi_j = 1$$

$$\int_{O} b_i(O_t) dO = 1$$

If the observations are discrete then the last integration will be a summation.

6. IMPLEMENTATION

The speaker independent speech recognition is implemented by training the system each 100 samples from different speakers consisting of 2000 words each. A test samples taken from a different set of 50 speakers each uttering 1000 words. All the samples were of British English and taken from TIMIT database. A partial output is given below.



Speech Project - Uma Maheswari

Text to create input sound file for speech Recognition

In this research, speech inputs are recognized by the system and executed. The project consists of two phases. The first part deals with recognizing the phonemes associated with the input voice using probabilistic neural network and the second phase involves pattern classification and displaying the input voice from the recognized phonemes using hidden markov model.

Recognition result from input sound file

In this research speech input are recognized by the system and executed the project consists of two phases the first part deals with recognizing the phony officials stated with the include boys using probabilistic neural network in the second phase involves pattern classification and displaying the input voice from the recognized phony infusion him Argos model

Recognition done

CONCLUSION

Speech recognition has a big potential in becoming an important factor of interaction between human and computer in the near future. A system has been proposed to combine the advantages of ANN's and HMM's for speaker independent speech recognition. The parameters of the ANN and HMM subsystems can influence each other. Encouraged by the results of the above described experiment, which indicate that global optimization of a hybrid ANN-HMM system gives some significant performance benefits. We have seen how such a hybrid system could integrate multiple ANN modules, which may be recurrent. A Neural Network with trained delays and widths and random weights classifies 98% of the phonemes correctly. A further refined speech recognition system can improve the accuracy to near 100%.

REFERENCES

[1] Medser L. R. and Jain L. C., "Recurrent Neural Network: Design and Applications." London, New York: CRC Press LLC,2001.
[2]D.A.Reynolds,"An Overview of Automatic Speaker Recognition Technology", Proc. ICASSP 2002, Orlando, Florida, pp. 300-304.
[3]Philip D. Wasserman, "Advanced methods in neural computing", Von Nostrand Renhold,1993
[4] D.F. Specht, "Probabilistic neural networks",Neural Networks, vol.3,pp. 109-118,1990.
[5] R.O.Duda, P.E. Hart, and D.G.Stork, "pattern classification", John Wiley and sons,second edition,2001.
[6] L.Deng and D.O'Shaughnessy, "Speech processing: a dyanamic and optimization-oriented approach", Marcel Dekker Inc.,2003
[7] R.Kronland-Martinet, J.Morlet and A.Grossman, "Analysis of sound patterns through wavelet transformation", International Journal of Pattern Recognition and Artificial Intelligence, Vol.1(2)
[8] John Coleman, "Introducing speech and language processing", Cambridge university press,2005
[9] L.Mesbahi and A.Benyetto,"Continuous speech recognition by adaptive temporal radial basis function," in IEEE international conference on systems,Man and Cybernetics,2004,pp.574-579
[10] H. Sakoe and S. Chiba, "Dynamic programming optimization for spokenword recognition", Proceedings of ICASSP-78, vol. 26, no. 1, pp. 43-49, 1997.
[11] "Timit acoustic-phonetic continuous speech corpus", National Institute of Standards and Technology

speech Disc 1-1.1, October 1990.

[12]Bourlard, H., Kamp, Y., Ney, H., and Wellekens, C. J. "Speaker-Dependent Connected Speech Recognition Via Dy- namic Programming and Statistical Methods," in *Speech and Speaker Recognition*, ed. M. R. Schroeder, Basel, Switzerland: Karger, pp. 115-148,1985.

[13] F. Jelinek ,"Statistical Methods for Speech Recognition", The MIT Press, 1998.

[14] L. R. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition", Prentice Hall,
Englewood Cliffs, New Jersey, 1993.

[15] E. Harborg, "Hidden Markov Models Applied to Automatic Speech Recognition", PhD Thesis, Norwegian Institute of Technology, Trondheim, Aug. 1990.

[16] Abbas Ahmadi,Fakhri Karray,Mohamed Kamel,"Modular-based classifier for phoneme recognition",IEEE International Symposium on Signal Processing and Information Technology,2006

[17] F. Karray and C. de Silva, "Soft computing and intelligent systems design: Theory, Tools and Applications", Addison Wesley Publishing, 2004.