IBN AL- HAITHAM J. FOR PURE & APPL. SCI. VOL.19 (2) 2006

Speaker Verification Using Hybrid Scheme for Arabic Speech

H.R.Mohammed
Department of Computer, College of Al- Education, University of Kufa

Abstract

In this work , a hybrid scheme for Arabic speech for the recognition of the speaker verification is presented . The scheme is hybrid as utilizes the traditional digital signal processing and neural network . Kohonen neural network has been used as a recognizer for speaker verification after extract spectral features from an acoustic signal by Fast Fourier Transformation Algorithm(FFT) .

The system was implemented using a PENTIUM processor , 1000 MHZ compatible and MS-dos 6.2 .

Introduction

Speaker recognition is the process of automatically recognizing who is speaking by using speaker specific information included in speech waves .(1)

Speaker recognition can be classified into speaker Identification and speaker Verification(2) . Speaker identification is the process of determining from which of the registeres speaker a given utterance comes.

Speaker verification is the process of accepting or rejecting the identity claim of a speaker in identification , the number of decision alternatives is equal to the size of the population where as in Verification there are only two choices accept or reject , rejard less of the population size(3) .

Therefore , speaker identification performance decreases as the size of the population increases , where as speaker verification performance approaches a constant , independence of size of the population , unless the distribution of physical characteristics of the speaker is extremely biased .

In the speaker verification , an identity claim is made by an unknown speaker , and an utterance of this unknown speaker is compared with the mode for the speaker whose identity is claimed . If the match , is good enough that is above a threshold , the identity claim is accepted .

A high threshold maker difficult for impostors to be accepted by the system , but at the risk of False rejecting Valid Users . Conversely , a Low threshold enable Valid Users to accepted consistenly , but at the risk of accepting impotors .(4)

The effectiveness of speaker Verification system can be evaluated by using the Receiver Operating Characteristics (ROC) Curve adopted of psychophysics .

The Roc curve is obtained by assigning two probabilities , the probability of correct acceptance and the probability of correct acceptance , to the vertical and horizontal axes respectively , and varying the decision threshold .

The Equal – Error Rate (EER) is a commonly accepted over – all measure of system performing .

If corresponds the threshold at which the False acceptance rste is equal to the false rejection rate . The Roc curves are shown in figure (1) .

The Arabic Phonemes

The ability to use it as a man – machine communication language seems to be a priority for the Arabic world . It has a number of unique characteristics that are not found in any other languages . Most important of this is that , it is often pronounced in an exact manner to its writing . Pronunciation rules are precisely defined (5) .

Spoken Arabic contains 29 consonants . Its consonants are identified by its factors :

1- Place of production (bilabial , labidental , dental, alveo-dental, palatal , velar , unvalued , pharyngeal, or glottal) .
2- Manner of production(op-plosive , fricative , affricate , nasal , glide , or semivowels)
3- Quality of voicing (emphatic or nonemphatic, and voiced or unvoiced)

The syllable in Arabic is based on the constructive component that is contained in its structure . The successive constructive element within a syllable boundary and made up of segmental phonemes of the language .

Each syllable has a main part that stands out and has prominence this part is refered to as the "nucleus " of the syllable . The remaining components are referred to as "margined factors " Acoustically , the nucleus is represented by formant structures and has more intensity thane than the marginal (6).

Neural Network

Neural network has been defind by Hecht- Nielson (2) as parallel distributed information processing structure consisting processing element interconnected unidirectional signal channels called connections . Each processing element has a signal output connection that branches ("fans out ") into as man collateral connection as desired ; each carries the same signal – the processing element output signal .

The processing element output single can be of any mathematical type desired . The information processing that goes on within each processing element can be defined arbitrarily with the restiction that it must be completely local ; that is , it must defend on the current values of the input signal arriving at processing element via impinging connection , and on values stored in processing element local memory .(7)

In mathematical terms (8), a neural network model has been defined as a directed graph with the following properties :

1- A state variable n_i is associated with each node I.
2- A real – valued weight w_{ik} is associated with each link (I_k) between two nodes I and K .
3- A real – valued θ is associated with each node I .
4- A transfer function f [n_k ,w_{ik} , θ ($k \neq I$) is defined , for each node I , which determines the state of the node as a function of bias ,of the weights of its incoming links , and of state of nodes connected to it by these links .

There are there common types of transfer function (5) (Activation function). Figure (2) shows the types of transfer functions we need to use.

Multi– layer perceptron (9) are feedforward nets with one or more layer of nodes between the input and output nodes. These additional layers contain hidden units or nodes that are not directly connected to both the input and output nodes. Below is a sample for a two–stage network structure :

Multi- layer perceptrons overcome many of the limitations of single – layers perpceptrons . The sigmoid function (2) has proven to be suitable for the transfer function .

$F(x) =(1/\exp(c.x))$ .....[1]

As it provides the pre – requisite for a suitable learning procedure because of its differentiability .

To use mutlilayered network (9) efficiently , one needs a method to determine their synaptic efficacies and threshold potentials A very successful method , usually called error back propagation , was developed e_m is the input and a_n is the output . If they are binary , they are represented by 0 or 1 .

The perceptron training algorithm (2) is a form of a supervised learn where the weights are adjusted to reduce errors whenever the network output does not match a known training target output .

The learning algorithm (5) for the perceptron is :

-Select random number for weights W_{kj} .

-Specify a random input vector ei. Let the selected vector be briefly described as $e = (e_1,e_2,e_3,.....,e_n)$ where by e_n=1 or 0 .

-Change the weights through

$W^{new}_{ij} = W^{old}_{ij} +\Delta W_{ij}$

With $\Delta W_{ij} = (\alpha .e_j . \varepsilon_I)$

Here ε , is the threshold the difference between the end output and the actual output at place of I, $\alpha>0$ is a random number , it is the learning rate , and is the tranfunction .

-Continue with (2)

-When the network produces the correct final vector for all input vectors then the algorithm ends(stops) .

Where k is over all nodes in the layers above node j . Internal node threshold are adapted in a similar manner by assuming that they are connection weights on links from auxiliary constant – valued inputs .

Convergence is sometimes faster if a momentum term is added and weight changes are smoothed by .

Unfortunately, there is no general criterion how the gradient parameter and the momentum parameter α are chosen . The optimal values depend on the problem to be learned .

The accuracy is found by (number of correct classifications) / (total number of patterns).

Linear Prediction Formulation:

The term LP refers to a variety of essentially equivalent formulations of the problem of modeling speech waveform . The difference among these formulations are often those of philosophy or way of viewing the problem .

The formulations are :-

1- Auto correlation method
2- Covariance method
3- Lattice method
4- Inverse filmier method
5- Spectral estimation method .
6- Maximum likelihood method .
7- Inner product method .

We now examin the first two methods .The third one is related to speech synthesis , and all other formulations are equivalent to one of the previous three formulations(9) .

Autocorrelation Method :

The key role in the liner prediction analysis is the correlation between diffcrent samples of the signal, such correlation can be computed in the process of determining LP(parameters that give an information about the formant frequecies and their bandwidths . This method assumes that the speech segment , is identically zero outside the interval 0<=m<=N-1

The speech signal was expressed as follows :

$S_n(m)=S(m+n)\ w(m)$ .....[2]

Where w(n) is a finite length window (e.g . Hamming window) . Clear lyaf $S_n(m)$ is non zero only for the period (0<=m<=N-1) there then the corresponding error $e_n(m)$, for a p′ th order predictor will be non zero over the interval (0<=m<=N+p-1) and can be defined as follows :

It can be seen that the prediction error is likely to large at the a beginning of the interval (specifically 0<=m<=p-1) because we are trying to predict the signal from samples that have arbitrarily been set to zero Likewise the error can be large at the end of interval (specifically N<=m<=N+p-1) because we are trying predict zero from samples that are none zero . For this reason , a window which tapers the segment , $S_n(m)$, zero is generally used .

The values of a_k that minimize E_n are found by setting for k=1,2,3,.....,p , this yields p linear equation .
The set of equation given in [1] can be expressed in P*P to plitz matrix of autocorrelation values as follows (6):-

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & . & R_n(p-1) & a1 \\ R_n(1) & R_n(0) & R_n(1) & . & . & a2 \\ R_n(2) & R_n(1) & R_n(0) & . & . & a3 \\ & & & & & \\ R_n(p-1) & R_n(p-2) & & & R_{n(}(0) & a_p \end{bmatrix} \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \\ R_n(p) \end{bmatrix}$$

Genera1 system for speaker Verification :

The experiment involve 1000 phrase Length Utterances of 30 speakers . An average misclassification rate of one percent with a "no decision " rate of four percent is obtained . Other experiments indicate that the utterances used for training purposes should preferably be collected over a relalively Long period of time .
There are general stages needed to implement the speaker recognition system . Figure (4) shows the general system structure using linear prediction analysis . As shown the test utterance must be first manipulated by an ends point detector , then applying linear prediction analysis, these two operations represent the preprocessing stage.

From the smoothed part (linear prediction coefficients)and the residue part (prediction residual signal) , it must produce a feature vector, this represents the feature extraction stage. Finally , performing acomparison with reference patterns via the matching or classification stage .

Short Domain Analysis : Time domain analysis transforms a speech signal into one or parameter signal , which usually varies much more slowly than the orig signal .

Speech From Silence Detect (End Points Detection): The speech signal was divided into blocks of (1103 samples)each . Then Zero Crossing Rate (ZCR) is calculated for each block in which this rate is high for speech samples and low for silence samples .We find speech samples only , after applying the following novel procedure which is designed and implemented to determine speech samples from silence samples :

```
Procedure silence samples;
FOR I=1 to BLOCKS DO   Zc[I]=0 ;
FOR I=1 TO BLOCKS DO  BEGIN ZEROC:=0 ;
                      FOR I=1 to SAMNO DO BEGIN inc
(L,1)
                      IF      ((Amp^[L]>      128)      and
(Amp^[L+1]<127))
                      OR((Amp^[L]<127))and
(Amp^[L+1]>128))
                      Them inc(ZEROC,1);end;
                      ZC[k]:=ZEROC ; end ;
L:=1;I:=1; COUNT:=1 ;
REPEAT
   IFZC[I]<>0 THEN
         BEGIN
           WORD:=I;INC(I,1);
           WHILE TEMP[I]<>0 DO
                 BEGIN
                      INC(COUNT,1);INC(I,1);
                 END;
            A[L]. AD :=WORD;A[L].CO:=COUNT;
             INC(L,1);
          END
    ELSE            INC(I,1);
       COUNT:=1;
    UNTIL I>=K;
REPEAT
    SK:=(A[I].AD-1)*SAMNO;
    SIZE :=SAMNO*A[I].CO;R:=1;
 FOR M:=SK TO SIZE +SK DO
               BEGIN
                 U[R]:=AMP^[M];INC(R,1);
               END;
    BLOCKWRITE(FK, U[1], SIZE,K);INC(I,1);
 UNTIL I>=L;
```

After this stage , a Linear prediction analysis could be performed on the speech data we fine the 14 th Linear prediction coefficients using autocorrelation method for speaker Verification processing .

Verification: In this stage, we used a neural network as a speech recognizer and we selected neural network .

Self-Organization Map (SOM): This network has been designed according to the following two neurobiologic observations : each cell responds to one specific input stimulus, and lateral interaction between neurons must cause the groupe of neurons around the neuron must be highly activated by the stimulus to be partially active .

SOM has two layers , an input layer and a Kohonen output layer This is shown in fig. (5) .The input layer is of a size determined by the user and must match the size of each pattern in the set of input patterns.

Each cell becomes an n-dimensional input X and computes its corresponding excitation values . The n- dimensional weight vectors W_j , j=1,2,3,.....,m , used for computation . Each cell learns to specialize on different regions of input space. When an input from such a region is fed into the network, the corresponding cell should compute the maximum excitation .

Kohonen proposed an interesting means of associating particular neurons with different input classes : he suggested that a particular input vector can be regarded as the image of neuron j in thc vector space defined by the set of input signals.

A Kohonen cell computes the Euclidean distance between an input X and its weight sector W.

learning Algorithm:

Kohonen learing algorithm uses a neighborhood function Φ , whose value $\Phi(j,k)$ represents the strength of the coupling between cell j and cell k during the training process . A simplc choice is defining $\Phi(j,k)=1$ for all cells j in a neighborhood of radius r of cell k and $\Phi(j,k)=0$ for all other cells . The learning algorithm for SOM is the following :

1)Start:The n-dimentional weight vectors {wj}, j=1,2,3,.....,m of the m computing units are selected at random . An initial radius N_c , a learning constant and neighborhood function are selected .

2) Step1 : Select an input vector Xand compute the distance d_j ($I \varepsilon N_c$)

$$d_i = \| X(n)-W_i(n)\| \qquad .....[3]$$

$$=\sum \sqrt{(x(n)-w_k(n))^2}$$

The distance is a measure for the mathc of X(n) with the weight vector $W_i(n)$, which may take another form of measure such as inner product of two vectors .

3) Step2: Find the central neuron j that satisfies

$$\| X(n)-W_j(n) \| = MIN \| X(n) -W_i(n) \| \quad \ldots\ldots\ldots\ldots [4]$$

, I=1,2,3,.....,m

The central neuron j is also called the winninn neuron whose weight vector is the best match with the input vector X(n) .

4)Step 3 : The weight vectors are updated using the neighborhood function and the update rule :

$$W_i(n=1) \leftarrow \begin{cases} W_i(n) + \acute{\alpha}\, \Phi\, (k,j)\, (X(n)-W_i(n)), & I \ \varepsilon \ Nc \\ W_i(n) & \text{otherwise} \end{cases} \quad \ldots\ldots [5]$$

5)Step 4: Iterate the computation by presenting a new input vector and returning to step 1, until the weight vectors stabilize their values .

After training the neural network , we noticed that most of the output layer cells , each one responds to only one phoneme and few of them (about 4 cells) was responding to more than one phoneme (2 or 3 phonemes). When we test the network (as a recognizer) by a pattern to identify the phoneme associated with it , we found that , if one of these few cells respond to this pattern , it is difficult to identify which phoneme from those is the correct one . The situation becomes more difficult when the phonemes that a word (pattern) contains activate two or three of these few cells . The space of choices will increased , so we have to find a way to eliminate this space to minimum .

Segmentation:

The most important and difficult stage in speech processing is how to identify the boundaries of word and the end point of phoneme and the starting point of the next phoneme along a word . Our segmentation algorithm is proved to be good and easy in performance in separating expected phonemes along isolated Arabic words .

It is based on a well defined mathematical formula depending on a few features . By this algorithm , we had clear boundaries for most

Arabic phonemes . The algo rithm fails in separating some Arabic consonant like (ع) and (ج) from vowels which represents the Arabic syllable (جي) ; you can see that the two phonemes are approximately closed because these two phonemes are nearly produced at the same region. We traced felt that the segmentation error to :

The highest formants power contents in such phonemer are very closed beccause they are produced from the Same region in the vocal tract .

Notice that people in south of Iraq and in Arabian Gulf prounce "ج"as"ي"

Recognition

In the recognition stage , we used the Self-Organization Map (SOM) which is characterized with simplicity . The training set consists of vectors which represent the Arabic phonemes .

We can not obtain a unique representation for each phoneme because each phoneme is affected by its neighboring phonemes and environment conditions . Our recognizer (SOM)faced a phenomenon in which a few of its output layer cells respond to more than one (2-3) phoneme , and in this situation the choice tree becomes bigger and the correct word (or syllable) is one of this tree branch and it is difficult to select it . This phenomenon reduced recognition accuracy . Also , the accuracy of recognition process is affected by the error of the previous stage (segmentation). Anyhow, though we faced these problems in recognition stage we get a good recognition rate .

Conclusions

FFT method has been used to compute frequency power along each frame (128 samples). Each spoken pattern consists of a variable number of frames . Kohonen Neural Network has been used as a recognizer for speaker verification after extract spectral features from an acoustic signal.

References

1.Bourland,H.(1996), Signal processingVIII: Theories and applications, Towards subbands – based speech recognition, Prentic-Hall,A380:365.

2. Davalo, E . Rawsthome, P. (1991),Neural networks, Macmillan,Fland.
3. Hopfield,J.(1991). Speaker- independent digit recognition using a neural network with time – delayed connection,IEEE ASSP,Magazine, 8 :(3) , 12-24, Sep .
4. Johnston,R. D.(1996),Are speech recognition still 98% accurate has the time come to repeal "Hyde ' s Law "? , IEEE ASSP,Magazine, 14 : 1 , Jan .
5.Kohonen,T.(1990),Digital signal processing with computer applications, 78,No.0,Sep .
6.Mandura,M.(1984),Synthesis of Arabic speech using phoneme-based synthesis , King Saud University , 10 (1,2), Saudi Arabia.
7.Rabinar, L.R., Juang ,B.(1993),Fundamentals of speech recognition, Prentic-Hall,London,A620:277.
8.Rabinar,L.R.,Juang,B.(1993),Fundamentals of speech recognition, Prentic-Hall,London,A630:277.
9.Robinson,T., Hochberg,M.(1996),Advanced topics in automatic speech and speaker recognition , IEEE ASSP,Magazine, 9 , No 13 , Jan .
10.Rojas,R.(1996),Neural network :Asystematic introduction,Springer, IEEE ASSP,Magazine, 18 , No 1 , july .

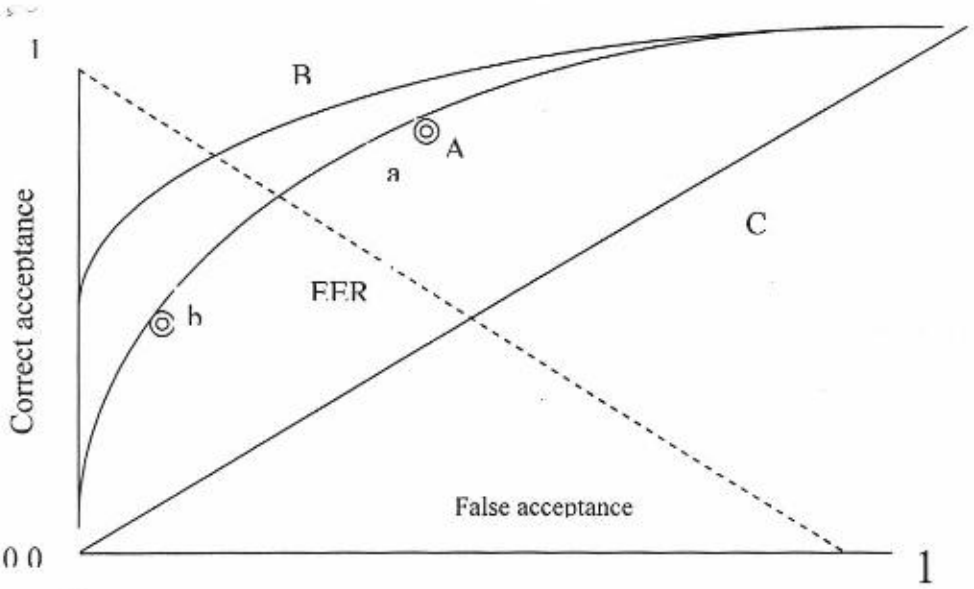

Fig. (1): Recceiver operating characteristic(ROC)curve (perforformance examples of three speaker Verification system A, B, AND C)

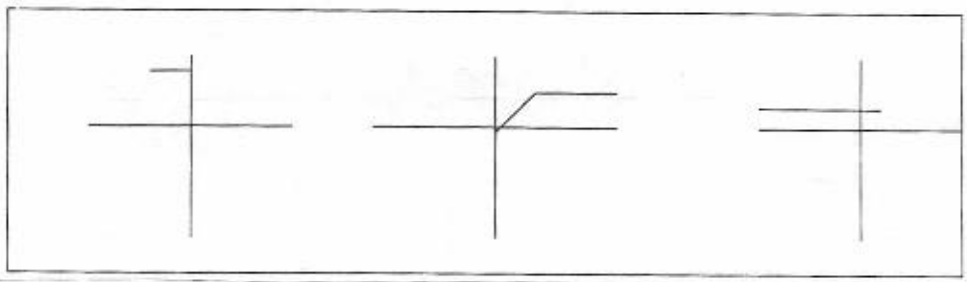

Fig. (2):types of transfer function

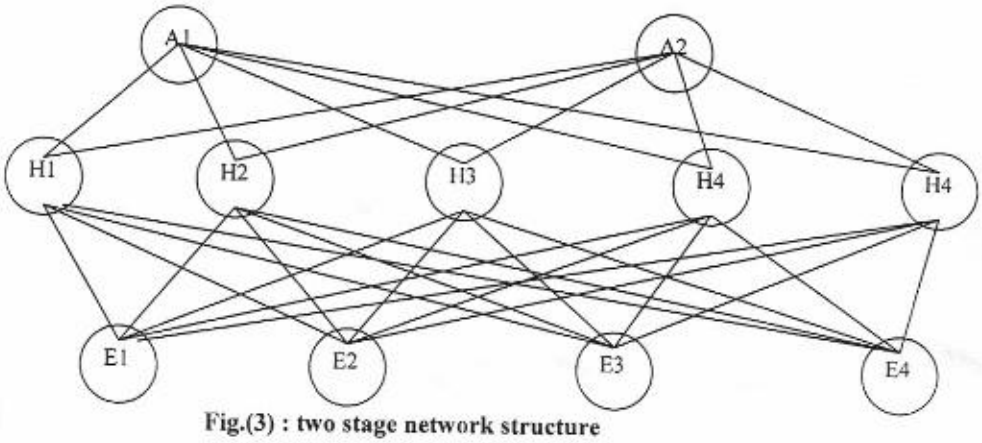

Fig.(3) : two stage network structure

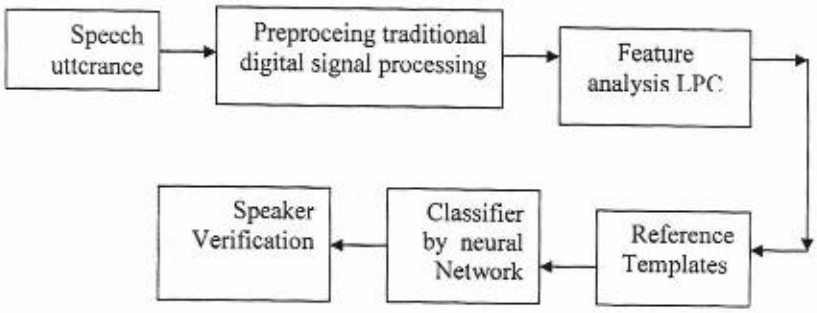

Fig. (4): General system for speaker Verification

مجلة ابن الهيثم للعلوم الصرفة والتطبيقية المجلد 19 (2) 2006

شرعية المتكلم باستخدام قاعدة هجينة للكلام العربي

هند رستم محمد

قسم الحاسبات، كلية التربية للبنات، جامعة الكوفة

الخلاصة

تم في هذا البحث تقديم اسلوب هجين لتمييز شرعية المتكلم ويعد هذا الأسلوب هجينا بسبب استخدام الطرائق التقليدية لمعالجة الاشارات الرقمية مع الشبكات العصبية . استخدمت الشبكة العصبية المسماة شبكة كوهونين(Kohonen) كمميز لشرعية المتكلم بعد استخلاص الخواص الطيفية التي تحويها الاشارة الصوتية باستخدام خوارزمية فورير السريعة(FFT).

طبق النظام باستخدام معالج نوع بنتيوم ذو سرعة 1000 ميكا هرتز واستخدم نظام التشغيل MS-dos 6.2 . MS-dos 6.2