Speaker Identification from Voice Using Neural Networks

Bijita Biswas* and Amit Konar
Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata 700 032

Received: 18 February 2002; accepted: 02 May 2002

The paper provides three different schemes for speaker identification of personnel from their voice using artificial neural networks. The first scheme recognizes speakers by employing the classical back-propagation algorithm pre-trained with known voice samples of the persons. The second scheme provides a framework for classifying the known training samples of the voice features using a hierarchical architecture realized with a self-organizing feature map neural net. The first scheme is highly robust as it is capable of identifying the personnel from their noisy voice samples, but because of its excessive training time it has limited applications for a large voice database. The second scheme though not so robust like the former, however, can classify an unknown voice sample to its nearest class. The time needed for classification by the first scheme is always unique irrespective of the voice sample. It is proportional to the number of feed-forward layers in the network. The time-requirement of the second classification scheme, however, is not free from the voice features and is proportional to the number of 2-D arrays traversed by the algorithm on the hierarchical structure. The third scheme is highly robust and mis-classification is as low as 0.2 per cent. The third scheme combines the composite benefits of a radial basis function neural net and back-propagation trained neural net.

1 Introduction

Several techniques on personnel identification from their voice are available in the literature on phonetics[1-3]. Most of these techniques discriminate the speakers from the pitch and the formant of the respective voice samples. Many of these methods are dependent on the voice contents or its specific format such as the CVC sequence (Consonant followed by a Vowel followed by a Consonant) and thus is not generic in its true sense. When the subjects are hostile to the environments, collection of such well-formatted speech sequence is difficult. For instance, one typical application of the speaker identification is in criminal investigation, where the subjects may not at all be harmonious to the scientific methods adopted for the system. The paper presents a novel scheme for speaker identification from the generic unformatted voice samples of the personnel and thus is unique in all respects. Soft computing tools, such as fuzzy logic and neuro-computing are gaining their importance in pattern recognition. Speaker identification from the voice of the subjects also belongs to this category of problems, and thus most of the available tools of soft computing may be utilized to handle this problem. Here three distinct models of artificial neural nets, such as the back-propagation (BP) algorithm and the self-organizing feature map (SOFM) neural nets and the radial basis function (RBF) neural net and BP combination will be employed to handle the present pattern recognition problem.

The work also compares the performance of the two neural techniques mentioned above with respect to the application of the neural nets[4-6] in the specific problem of speaker identification. An analysis of the comparison reveals that the BP algorithm[7-9] with momentum term in the gradient descent learning expression has a good convergence to the global minima of the error surface and thus is amenable for speaker identification even with small Gaussian noise around the measured input voice instance. One practical difficulty faced in the experimentation phase with the back-propagation algorithm is the excessively large training time. In fact the training time is higher order polynomial of the number of sample input-output instances. In our case the training input instance is the features of the voice samples, and the output instance is the speaker number. The experimental results also demonstrate that when the training samples exceed 700 instances the training...
time is of the order of several days for arriving at a RMS error margin of 0.001 or lower.

The alternative approach of classifying the input training samples by a topology sensitive adaptation scheme fortunately requires only a few minutes to handle the problem of the same dimension mentioned above, and thus, of course, draws much attention. The SOFM based scheme has experimentally been found to have a relatively poor classification accuracy than the BP based scheme. Failure in classification takes place when the voice samples of two or more patterns do not have significant difference. In our implementation the percent error in classification has been found to be significantly low with respect to the works already done with SOFM. This could have been realized because of one specialized feature pitch, defined differently that can isolate significant spatial difference in the topological clustering process on the 2-D planes.

Formal definitions of speech features to be used for speaker identification are given in Section 2. A scheme for speaker identification using the classical BP algorithm is presented in Section 3. The SOFM based scheme for speaker classification is presented in Section 4. The RBF-BP scheme for speaker recognition is introduced in Section 5. The experimental analysis for optimum choice of features for speaker recognition is presented in Section 6. The conclusions will be listed in Section 7.

2 Voice Feature Extraction

Voice samples from 500 speakers within a given locality in their native language have been collected and processed by a digital sonogram (Model No. 5500 maintained by Indian Statistical Institute, Calcutta). The sonogram samples the voice signal at a fixed rate of 10240 Hz and the samples are quantized to binary numbers with a truncation noise of 1/256 the dynamic range of the signal or less. The voice features have then been directly analyzed by the sonogram that provides the first few formants \( F_1, F_2, \ldots, F_5 \), the pitch of the voice signal defined below and the maximum power spectrum of the voice signal. The power spectrum at the selected sample frequencies is also treated as the voice features. The following definitions are needed to explain the subsequent part of the paper.

**Definition 1** — A **formant** is defined as average of the fundamental (harmonic) frequencies obtained in a voice spectrum over a given interval of time. Formally, if \( f_i \) is the fundamental (harmonic) frequency of the voice at time \( i \), then the \( j \)-th formant \( F_j \) is defined as:

\[
F_j = \frac{1}{T} \sum_{i=0}^{T} f_i / T \quad \ldots \quad (1)
\]

Usually the first formant \( F_1 \) corresponds to the fundamental frequency, while the second, third, \( n \)-th formant refers to the corresponding harmonic signal present in the voice. The averaging in the definition of the formant is needed from the computational standpoint as the fundamental formants have a small deviation in the speech sample. Figure 1 explains the frequency deviation in the respective fundamental and the harmonic signals of a voice spectrum.

**Definition 2** — **Power spectral density** of a voice signal is defined as the instantaneous root mean square power of a voice signal spread over the available frequency range. It is usually denoted by \( p^2(f) \).

For the computation of the total power in a frequency interval \( f_1 \) to \( f_2 \), we usually need to integrate \( [p^2(f)] df \) within the said bounds of \( f_1 \) and \( f_2 \).

**Definition 3** — The **peak power spectral density** is defined as the peak amplitude(s) of the power spectrum within a pre-defined frequency interval of \( f_1 \) to \( f_2 \). In case the power spectrum has several peaks, they are sorted according to their amplitudes in a descending order and the first three elements of the sorted array are used as the voice feature.
Definition 4 — The pitch is a measure of the large amplitude changes in a signal within a small time-span. For convenience of measurements, a voice signal is first sampled at a high rate, typically of the order of 10 kHz, and then quantized. A count of the difference between the successive quantized levels above a predefined threshold over a time duration of 1 s is absolutely expressed as the measure of pitch (Figure 2).

The voice signal after digitization can also be analyzed in a MATLAB environment for the evaluation of the above features. The estimated values thus obtained may directly be ported to the neural net toolbox under MATLAB, or can alternatively be realized with the user’s own C-codes for neural networks. In the present context we first restricted our scheme for speaker identification using MATLAB and then generated our own C-codes for neural net stimulation for on-time speaker recognition.

3 Speaker Identification Using The Back-Propagation Algorithm

3.1 The Training Cycle

The neural network model with four layers having ten neurons in the input layer, twelve neurons in each hidden layer and nine neurons in the output layer have been used for the computer simulation. It may be noted that the output of the neural net here generates an encoded output, which needs to be decoded externally to identify the speaker. This was done so as to reduce the number of output lines even for a speaker set of 500.

In the simulation under MATLAB the network has been configured to use the LOGSIGMOID function, which generates the output between 0 and 1. The training has been done by TRAINBPX, which uses both the momentum and the adaptive learning rate to speed up the learning. The learning rate and the number of iterations (epoch) in the present context have been selected judiciously so as to the minimize the sum-square error criterion accurately. The TRAINBP algorithm provides an autonomous adaptation of the learning rate.

The TRAINBP algorithm allows a forward pass through the network and subsequently determines the error at the nodes of the output layer by comparing them with the supplied target values. The error thus generated at the output layer is then propagated back following the steps of the back-propagation algorithm. Consequently, the weights of the nodes preceding to the layer where error has been back-propagated, are adapted by the gradient descent learning principle. The process is repeated from the output layer to the last hidden layer, then from the last hidden layer to the first hidden layer and finally to the input layer. The whole sequence of weight adaptation steps together is called one learning epoch. Usually the learning rate and the parameters a and b used in the weight adaptation expressions are kept fixed within a learning epoch.

Thus, at the end of a learning epoch, error at the output layered nodes are calculated. If the new error exceeds the old error by more than a predefined ratio (typically 1.04), new weights and error are discarded, and the learning rate is decreased (typically by multiplying by 0.7); otherwise, the new parameters are saved. If the new error is less than the old error, the learning rate is increased (typically by multiplying by 1.05).

3.2 The Recognition Cycle

Once the neural net is trained it may be used for recognition of the speaker from their voice. In case of the back-propagation learning the recognition process is very fast; it only needs to have a forward pass over the network. The outcome at the output layer thus obtained is then decoded to get the exact speaker number. In case the outputs are close to but less than 1, we consider it as 1. Similarly, if one or more
outputs of a given network is close to but not equal to zero, we call it logic 0.

3.3 Results

Experiments with a large number of voice database reveal that the BP algorithm can safely recognize more than 98 per cent of the voice data. The remaining 2 per cent of the voice data could not be recognized for their poor and ambiguous measurements. The training time of the algorithm is very high, of the order of several hours. An alternative neural topology thus may be used to solve the same problem.

4 Speaker Identification Using Self-organizing Feature Map

4.1 The Training Cycle

The second approach to solve the speaker identification problem is to employ a hierarchical self-organizing feature map (SOFM). The topology of a hierarchical SOFM is a tree where each node in the tree corresponds to one 2-D array where the input vectors are mapped.

The SOFM works using the following principle. The weights of all neurons in the 2-D planes are first randomized. The input feature vectors are then mapped onto the first 2-D plane of neurons by the following learning rule:

\[ W_{\text{new}} \leftarrow W_{\text{old}} + \lambda \left( I_k - W_{\text{old}} \right) , \ldots (2) \]

where \( W_{\text{old}} \) is the weight vector of a given neuron, \( W_{\text{new}} \) is the adapted value of the weight vector and the norm \( \| \cdot \| \) denotes the Euclidean distance between the two vectors. After each weight adaptation the old weight vector is replaced by the new vector.

The above process corresponds to mapping an input vector to the neuronal plane. The process thus may be repeated for all possible input training vector. In case, more than one input vector is mapped at the same neuron, they need to be re-mapped onto a new 2-D plane. The process is thus repeated by constructing newer planes until all neurons are mapped separately on individual nodes of the network. Thus the hierarchical organization of the tree is justified (Figure 3).

4.2 The Recognition Cycle

It may be mentioned here that during the training phase the input vectors used for training the network comprises all the components mentioned in BP algorithm plus one, the last one being the speaker number. This is needed because in the recognition phase when the unknown input vector is searched against the hierarchy of nodes, the first ten fields corresponding to the features only are compared. When a successful match is found on the first plane, it is checked whether more than one input vectors had been mapped there. If no, then the 11th field of the stored weight vector is the result to be submitted to the users. If yes, then we need to search the input vector in the next level of the tree. The search process thus continues until the leaf node in the tree is reached. Under this case, on completion of the search process, the last found node's 11th component is regarded as the speaker number.

5 Speaker Identification by Composite Use of RBF Neural Net and Back-propagation Algorithm

Back-propagation algorithm is a good technique for speaker identification from voice features of the speakers. But as the number of speaker increases the training time also increases to a great extent. For example, when the number of speakers increases from 50 to 500 the training time increases by a factor of 20 (approximately) for a prescribed level of mean square error sum at the output layer of the order of 0.001. Typically the training time for the back-propagation algorithm for approximately 500 speakers on an IBM Pentium type PC is around 40 h for a mean square error sum of the order of 0.001. Thus, back-propagation is not a good choice for large number of training instances. Because of the small training time of the SOFM or the Radial Basis Function (RBF) type neural nets, they are good choice for speaker identification. The SOFM and RBF networks usually classify the input instances into clusters. Thus one to one mapping from input to output instances is not amenable by these networks. A combination of an RBF or an SOFM network with a back-propagation type neural net in cascade seems to
be an ideal choice for speaker recognition. In this section, the scope of the composite use of the RBF and the back-propagation type neural net in the speaker identification problem is examined.

For our experiment we considered 100 speakers. An RBF type neural net comprising of two layers only (the input layer and the RBF layer) is employed to select one of ten three layered feed-forward neural net trained with back-propagation algorithm. Each of these ten neural nets is trained with ten input-output instances that are sufficiently close to form a cluster. The RBF layer of the radial basis function neural net comprises of ten neurons, each tuned to the cluster center of the ten close patterns that together form a cluster. Such training requires significantly small computational time, but the trained network can correctly identify the one out of 100 speakers at the output layer of the back-propagation neural net assembly (Figure 4).

The RBF function in the present context is a typical Gaussian function $G(\mu, \sigma^2)$, whose mean $\mu$ is the component-wise average of the patterns within a cluster, and variance $\sigma^2$ is the variance of the patterns themselves. Consequently, the cluster centers are sufficiently different to identify the class to which an unknown speaker belongs.
The mean and variance of the cluster centers for each node in the RBF layer are computed, prior to the application phase. Thus when a new training instance appears at the input layer of the RBF-type neural net, the network can instantaneously determine the class of the unknown speaker. The back-propagation neural nets are also pre-trained with the ten voice features as input and the speaker identification number as the output.

6 Optimal Choice of Features in Speaker Identification

Selection of features is an important consideration in speaker identification problem. A series of experiments have been carried out to determine the minimal but the best possible features suitable for the speaker recognition problem. The experiments reveal that the first few formant
Table 1— Accuracy of correct identification of speakers on feature selection

<table>
<thead>
<tr>
<th>Selected features</th>
<th>Percentage of speakers correctly identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$ $F_2$ $F_3$ $F_4$ $F_5$ Pitch TSP PPSD $ST_1$ $ST_2$</td>
<td>99.8</td>
</tr>
<tr>
<td>$F_1$ $F_2$ $F_3$ $F_4$ $F_5$ Pitch TSP PPSD $ST_1$ $ST_2$</td>
<td>99.7</td>
</tr>
<tr>
<td>$F_1$ $F_2$ $F_3$ $F_4$ $F_5$ $F_6$ Pitch TSP PPSD $ST_1$ $ST_2$</td>
<td>99.2</td>
</tr>
<tr>
<td>$F_1$ $F_2$ $F_3$ $F_4$ $F_5$ $F_6$ $F_7$ Pitch TSP PPSD $ST_1$ $ST_2$</td>
<td>95.0</td>
</tr>
<tr>
<td>$F_1$ $F_2$ $F_3$ $F_4$ $F_5$ $F_6$ $F_7$ $F_8$ Pitch TSP PPSD $ST_1$ $ST_2$</td>
<td>94.0</td>
</tr>
<tr>
<td>$F_1$ $F_2$ $F_3$ $F_4$ $F_5$ $F_6$ $F_7$ $F_8$ $F_9$ Pitch TSP PPSD $ST_1$ $ST_2$</td>
<td>90.0</td>
</tr>
<tr>
<td>$F_1$ $F_2$ $F_3$ $F_4$ $F_5$ $F_6$ $F_7$ $F_8$ $F_9$ Pitch TSP PPSD $ST_1$ $ST_2$</td>
<td>91.2</td>
</tr>
<tr>
<td>$F_1$ $F_2$ $F_3$ $F_4$ $F_5$ $F_6$ $F_7$ $F_8$ $F_9$ Pitch TSP PPSD $ST_1$ $ST_2$</td>
<td>66.3</td>
</tr>
</tbody>
</table>

Corresponding field of the input training instance by a small positive number = 0. 85. The other fields of the input instance are normalized by dividing their absolute value by the maximum value under the given field. Such normalization process prohibits saturation of the sigmoid function in the back-propagation algorithm.

7 Conclusions

The paper presented a new approach to speaker identification by using the composite benefits of a RBF neural net and a number of three-layered feed-forward neural nets trained with back-propagation algorithm. The RBF neural net classifies the input feature space of the speakers into clusters and subsequently selects the appropriate Back-propagation neural net for determining the speaker under a class.

Selection of features has been given main consideration in the speaker identification problem. Experimental results demonstrate that the speaking time vector, the first three formants, the peak power spectral density and the total power delivered in the speech signal together can identify the speaker to high accuracy, i.e. around 99 per cent.

References


