Hybrid Speech Recognition System based on Wavelet 9/7 and Mel-Frequency Cepstral Coefficient

Sozan Mahmood and Mihran Abdulrahim

Abstract—This research investigated the idea using a hybrid approach (uses wavelet transforms tap 9/7 and MFCC) as a feature extraction method to recognize speech signals. Then comparing this hybrid system with two other systems, first system is based on wavelet tap 9/7 and the second one is based on MFCC. A back-propagation neural network was utilized to recognize the speech segments. MFCC based system gave better recognition rate than the wavelet system. The recognition rate of the proposed hybrid system is give better result than the recognizers based on MFCC or wavelet.

Keywords— MFCC, Neural Network ,Signal processing, Wavelet transform.

I. INTRODUCTION

Speech recognition provides a very powerful element for user interfaces and in particular human-robot interaction. While commercial, state-of-the-art speech recognition technology may be useful for telephony applications.[1]

Previous research relied heavily on the Hidden Markov Model in Speaker recognition [2].

The most important parts of a speech recognition system are the feature extraction and the recognition methods. The feature extraction step converts the properties of the signal which are important for the pattern recognition task to a format that simplifies the distinction of the classes. The recognition step aims to estimate the general extension of the classes within feature space from a training set [3].

In this work, three speaker independent isolated speech recognition systems have been developed, back propagation neural network is used as the classifier to classify different words to their classes, then the command is transmitted to the remote device via serial output port.

Mainly, two different feature extraction methods in recognition process were followed. The MFCC based approaches was investigated, and wavelet transform is used also. The combination of these two approaches is implemented to design a hybrid approach (wavelet and MFCC). For the matching in these systems BPNN has been utilized.

Fig. 1 presents the layout of the neural based system. The lifecycle of the processing passes through two stages: training stage and recognition stage. In the first stage of the design, the speech is appropriately processed to be input to the neural networks. By this we imply feature extraction achieved through modeling the human vocal tract using wavelet 9/7 tap then converted to the more robust cepstral coefficients using MFCC. The second stage of the design is to train the system for different utterances of the words. These utterances should constitute a good sample set of the various conditions and situations in which the word may be pronounced[4].

II. SYSTEM LAYOUT

Three speaker independent isolated speech recognition systems have been developed, back propagation neural network is used as the classifier to classify different words to their classes, then the command is transmitted to the remote device via serial output port.

III. SPEECH PREPROCESSING

The basic idea behind speech processing is to produce more data reduction ability and easy to analysis. After the wave file is digitized and stored in memory, and before extracting features from the signal further steps are applied to the signal to remove unwanted parts from the signal, in order to get more accurate and better performance for the system. This stage consists of two steps, normalization and speech boundary detection.

The obtained signal after digitization will be passed through a normalization process as in (1), by dividing each sample by
the square root of the sum of square of all the samples divided by the number of samples in the signal.

\[
S[i] = \frac{S[i]}{\sqrt{\frac{\sum S[j]^2}{N}}}
\]

Where \( S[i] \) is the ith sample, \( N \) is the number of sample in the signal.

Since not all information included in a speech signal is relevant for recognition, some processing techniques are applied to discriminate relevant information and exclude. In this part of the proposed system, the features are extracted by using wavelet transform, and then the extracted data is taken as input to MFCC. The system was compared with results obtained from the wavelet and MFCC alone.

Wavelet analysis is a relatively new mathematical discipline, which has generated much interest in both theoretical and applied mathematics over the past decade. Wavelets have the ability to analyze different parts of a signal at different scales. The wavelet transform (WT) is a transformation that provides time-frequency representation of the signal\[5\].

The primitive lifting process steps for the 9/7 filter are described from the following equations. First the input sequence \( x_i \) are split into even and odd parts \( d_i^0 \) and \( s_i^0 \):

\[
d_i^0 = s_{2i+1},
\]

\[
s_i^0 = s_{2i},
\]

Then two lifting steps is performed on the two spitted sequences respectively and the outputs are denoted as \( s_i^\alpha \) and \( d_i^\alpha \) \((\alpha=1,2)\).

\[
d_i^1 = d_i^0 + \alpha \times (s_i^0 + s_{i+1}^0)
\]

\[
s_i^1 = s_i^0 + \beta \times (d_i^0 + d_{i+1}^0)
\]

\[
d_i^2 = d_i^1 + \gamma \times (s_i^1 + s_{i+1}^1)
\]

\[
s_i^2 = s_i^1 + \delta \times (d_i^1 + d_{i+1}^1)
\]

Finally thought the scaling factors \( K_2 \) and \( K_1 \), the low-pass and high-pass wavelet coefficient can be obtained as shown in Fig.2

\[
d_i = K_2 \times d_i^2
\]

\[
s_i = K_1 \times s_i^2
\]

Where \( K_2 = \zeta \) and \( K_1 = 1/K_2 \), and the value of \( \alpha, \beta, \gamma, \delta, \) and \( \zeta \) are the value of lifting coefficients.

After applying wavelet it is not necessarily to get the desired number of coefficients, and even number of samples may still too much for recognition, and it may take forever time for the neural network to learn. So an additional step “partitioning” should be added.

**Fig. 2** (a) the original speech signal, (b) the absolute value of signal, (c) first band filter, (d) second band filter, and (e) the last band filter.

In partitioning the signal is divided to a number of frames, where every frame has \( N \) samples, the features are calculated from the mean value for each frame.

The reason behind the popularity of MFCC based feature extraction is that it have been shown to mimic the way the human ear responds to acoustic input. MFCC based feature extraction is implemented through many steps. Each step is motivated by perceptual or computational considerations. Mel frequency scale is linearly space below 1 kHz and logarithmical above. As previous studies mentioned, number of triangular filters can vary between 20 and 40. In this work (20, 24, and 26) was taken as the number of triangular filters.
This is done by multiplying the power spectrum with each triangular filter in the filter bank and adding the result.

The following steps illustrate the whole MFCC calculation process:
1. Pass vector S (original signal) through a pre-emphasis filter.
2. Divide N samples to overlapping frames.
3. Calculate and apply hamming window on each frame.
4. Convert each frame to frequency domain.
5. Calculate power spectrum of vector.
6. Calculate Triangular filters and apply them to each frame.
7. Calculate triangular filters using equation.
8. Calculating the logarithm of the mel-scaled filter bank energies.
9. Apply (DCT) Discrete Cosine transform (DCT) of the mel-scaled log-filter bank energies to calculate MFCCs.

Many parameters involved in evaluating the system's performance. Various values were tested for each of the parameters.

The structure of the back propagation neural network consists of three layers: first layer has (20, 40, 50 and 60) input neurons which are fully connected to the hidden layer see TABLE I. The last layer is the output layer consisting of 4 neurons whose output uses to binary encoding. All three layers are fully feed forwarded trained.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>THE EFFECT OF INPUT NODES ON THE OVERALL RECOGNITION RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input nodes</td>
<td>Recognition rate %</td>
</tr>
<tr>
<td>Wavelet 9/7tap</td>
<td>Proposed system</td>
</tr>
<tr>
<td>20</td>
<td>54</td>
</tr>
<tr>
<td>40</td>
<td>58</td>
</tr>
<tr>
<td>50</td>
<td>57</td>
</tr>
<tr>
<td>60</td>
<td>55</td>
</tr>
</tbody>
</table>

Also the effects of learning rate value were tested. Since the value of learning rate and momentum have a great importance in learning process of the NN, different values for these two variables were taken varied from 0.01 to 0.95 during training see TABLE II. Giving learning rate a proper value has significant affect on the learning process performance and accuracy. There is no predefined specific value for this parameter; it is highly based on the problem. If the value of learning rate is too small the learning process will be too slow, and if its value is large, it may pass over weight values which lead to good learning.

After completing training the NN network can be used to classify the different uttered words to their different class. In this work a different numbers of hidden nodes have been tested to reach the best training results see TABLE III. Also the effect of number of triangular Mel filter banks on the recognition rate of the recognition system based on MFCC shown in TABLE IV.

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>THE EFFECT OF LEARNING RATE ON THE OVERALL RECOGNITION RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate</td>
<td>Recognition rate %</td>
</tr>
<tr>
<td>Wavelet 9/7tap</td>
<td>MFCC</td>
</tr>
<tr>
<td>0.01</td>
<td>59</td>
</tr>
<tr>
<td>0.05</td>
<td>52</td>
</tr>
<tr>
<td>0.1</td>
<td>51</td>
</tr>
</tbody>
</table>

In the first stage of the design, the speech is appropriately processed to be input to the neural networks. These utterances should constitute a good sample set of the various conditions and situations in which the word may be pronounced[1].

IV. EXPERIMENTAL RESULTS

In order to evaluate the performance of the speech recognition proposed system, a set of experiments was performed. For the purpose of testing the performance of the proposed speech recognition system, all utterances in the training data set is used, then the test is repeated on the set of the data that dedicated to the testing phase. To evaluate the performance of the proposed system, five words (in English language) are used. Ten speaker’s utterances. Each user repeats a statement ten times. The statement consists of uttering the five words with a small silent period between each word. Therefore the total collected utterances for each user are 50.

Neural networks attempt to mimic some or all of the characteristics of biological neurons that form the structural constituents of the brain [6]. In this paper, the back propagation neural network was adopted since it has been successfully applied to many pattern classification problems including speech recognition and our problem has been considered to be suitable.
V. CONCLUSION

The wavelet transforms tap 9/7 is lossy method and therefore it does not give a very good recognition rate. The wavelet coefficients led to a maximum recognition rate (66%), when 40 coefficients were taken as feature vector, number of hidden nodes 15, and learning rate 0.1. In this system the more coefficients gives less recognition rate, and more time required for training.

MFCC based system gave better recognition rate than the wavelet system. The maximum recognition rate (90%) was obtained when the number of utilized Mel filter banks was 40, frame size was 512, hidden nodes 30 and learning rate 0.01.

The proposed hybrid system based on wavelet transforms and MFCC is useful for speech recognition task. The wavelet transforms boosts the energy of the signal, and this signal goes through the MFCC feature extraction approach. The recognition rate for this system in comparison to the two previous ones is better. The maximum recognition rate is (96%), when frame size is 512, Mel filter banks are 40, learning rate is 0.05, and hidden nodes are 30.

ACKNOWLEDGEMENTS

I would like to acknowledge the anonymous reviewers for their comments to improve the quality of this paper.

REFERENCE


Dr. Sozan Abdulla was born in Sulaimani 1965, she received her B.Sc. and M.Sc from computer science at Mosul University. She received her PhD at Sulaimani University. Currently she is Prof. assistant at Sulaimani University. Her research interests include AI, signal processing, and image processing.

Mihran Abdulaim was born in Sulaimani 1988, she received her B.Sc. and M.Sc from computer science at Sulaimani University(2011). Currently she is lecturer at Sulaimani University. Her research interests include signal processing, and pattern recognition.