Speech Recognition using FIR Wiener Filter

Deepak ¹, Vikas Mittal ²

¹ Department of Electronics & Communication Engineering, Maharishi Markandeswar University, Mullana (Ambala), INDIA
² Department of Electronics & Communication Engineering, Maharishi Markandeswar University, Mullana (Ambala), INDIA

ABSTRACT

This paper presents a speech recognition system using cross-correlation and FIR Wiener Filter. The algorithm is designed to ask users to record words three times. The first and second recorded words are different words which will be used as the reference signals. The third recorded word is the same word as one of the first two recorded words. The recorded signals corresponding to these words are then used by the program based on cross-correlation and FIR Wiener Filter to perform speech recognition. The algorithm then give the judgment that which word is recorded at the third time compared with the first two reference words. The results showed that the designed system works well when the first two reference recordings and the third time recording are done by the same person. The designed system also reported errors in situations; the first two reference recordings and the third time recording are from different people. Thus, the designed system works well for the purpose of speech recognition.

Key words: Speech recognition, Cross-correlation, Wiener Filter & Auto-correlation.

I. INTRODUCTION

Speech recognition is a popular topic in today’s life because of its numerous applications. For example, consider the applications in the mobile phone in which instead of typing the name of the person who user want to call, the user can just directly speak the name of person to the mobile phone, and the mobile phone will automatically call that person. Another example is instead of typing the keyboard or operating the buttons for the system, using speech to control system is more convenient. It can also reduce the cost of the industry production at the same time [1-4]. Using the speech recognition system not only improves the efficiency of the daily life, but also makes people’s life more diversified. However, speech is a random phenomenon and thus its recognition is a very challenging task. Therefore, this paper investigates the correlation and FIR Wiener Filter based algorithms of speech recognition [5-6]. The recognition process comprises of three input speech words, in which two are reference speech words and the third is target speech word. The target speech word is compared with the two reference speech words and the efficiency of algorithm is evaluated in terms of percentage of the words recognised.

II. SYSTEM MODEL

In this paper, the system is designed to record user’s voice using the microphone and soundcard of a computer. To improve the quality and reduce the effect of DC level, the mean value of recorded signals is deducted from itself. To obtain a better quality of signals the sampling frequency is set as 16 kHz, as used in most non-telecommunications applications [6-7]. The length of the recorded signal is 2 seconds. After recording the signals, the same are analysed in frequency domain in this paper. For this purpose, the FFT is applied on three signals and the obtained spectrums are then normalised to an interval [0 1], so as to have the same scale suitable for comparison purposes. The next step is to use this information to do speech recognition. For this purpose, firstly cross-correlation function between the targeted signal and the two reference signals is computed.

A. Cross-correlation function

Speech is a random phenomenon, so even for the same speaker, the same words may have different frequency bands. This is due to the different vibrations of vocal cord. Thus, the shapes of frequency spectrum obtained may be different. But, the similarity between these spectrums determines the degree of recognition between the speech signals. This forms the bases of this paper for speech recognition [7-9]. More specifically, the speech recognition is done by comparing the spectrums of the third recorded signal and first two recorded reference signals. This can be done by computing the cross-correlation of two signals; which will give the shift parameter, also referred as frequency shift. The definition of the cross-correlation for two signals is as below:
From the above equation (eqn 1), two important information regarding the cross-correlation function can be obtained. Firstly, when two original signals have no time shift, their cross-correlation should be at its maximum value. Secondly, the position difference between the maximum value position and the middle point position of the cross-correlation function is the length of time shift for two original signals. Now, assuming the two recorded speech signals for the same word are totally the same, so the spectrums of these signals should be the same. This means the cross-correlation function of these signals will be totally symmetric. However in practise, the spectrum of recorded signals (even for the same word) is usually not same. This means their cross-correlation graph will not be symmetric. This symmetric property of cross-correlation function is used in this paper to do speech recognition. More specifically, by comparing the level of symmetric property for the cross-correlation, the system can make the decision that which two recorded signals have more similar spectrums and thus recognise each other.

Mathematically, if we set the frequency spectrum’s function as a function \( f(x) \), then, according to the axial symmetry property definition: for the function \( f(x) \), if \( x_i \) and \( x_2 \) are axis-symmetric about \( x=x_2 \), then \( f(x_1) = f(x_2) \). For the speech recognition comparison, after calculating the cross-correlation of two recorded frequency spectrums, there is a need to find the position of the maximum value of the cross-correlation and use the values right to the maximum value position to minus the values left to the maximum value position. The absolute value of this difference and the mean square-error is calculated. If two signals match better, then their cross-correlation is more symmetric. And if the cross-correlation is more symmetric, then the mean square-error should be smaller. By comparing this error, the system decides which reference word is recorded at the third time. However, this algorithm is more suitable for the noise free conditions, which is difficult to obtain in practise. Therefore, a FIR Wiener Filter is added to this to make it more robust. This is discussed below.

### B. The FIR Wiener Filter

The principle of FIR Wiener filter is shown in Figure 1 below. It is used to estimate the desired signal \( d(n) \) from the observation process \( x(n) \) to get the estimated signal \( \hat{d}(n) \). It is assumed that \( d(n) \) and \( x(n) \) are correlated and jointly wide-sense stationary.

![Figure 1: Wiener filter for speech recognition purposes.](image)

From Figure 1, assuming the filter coefficients are \( w(n) \), so the output \( \hat{d}(n) \) is the convolution of \( x(n) \) and \( w(n) \):

\[
\hat{d}(n) = w(n) * x(n) = \sum_{m=0}^{\infty} w(m) x(n-m)
\]

Then, the error of estimation is

\[
\epsilon(n) = d(n) - \hat{d}(n)
\]

\[
\epsilon(n) = d(n) - \sum_{m=0}^{\infty} w(m) x(n-m) = \sum_{m=0}^{\infty} w(m) x(n-m) - d(n)
\]

From equation 3, the purpose of Wiener Filter here is to choose the suitable filter order and find the filter coefficients with which the system can get the best estimation. In other words, with the proper coefficients the system can minimise the mean-square error:

\[
E = E[|\epsilon(n)|^2] = E[(d(n) - \hat{d}(n))^2]
\]

The minimization of the mean-square error is done using the suitable filter coefficients. This means the derivative of \( E \) with respect to \( w(n) \), as:

\[
\frac{dE}{dw(n)} = \frac{dE}{dw^2(n)} = \frac{dE}{dw(n)} E[\alpha(n)\epsilon(n)] = E[\alpha(n) \frac{d\alpha(n)}{dw(n)}] = 0
\]

Thus, from equations 3 and 5:

\[
\frac{d\alpha(n)}{dw(n)} = -\alpha(n) (n - k)
\]

So, the equation (5) becomes:

\[
\frac{dE}{dw(n)} = E[\alpha(n) \frac{d\alpha(n)}{dw(n)}] = -E[\alpha(n)\alpha(n) (n - k)] = 0
\]

Then, we get:
The above equation is known as orthogonality principle or the projection theorem [6]. Using equation (5), we have

\[ E\{e(n)x^* (n-k)\} = E\left\{ \left[ d(n) - \sum_{k=0}^{p-1} w(k)x(n-k) \right]x^*(n-k) \right\} = 0 \]  

Rearranging, the equation (9) to have:

\[ E\{d(n)x^*(n-k)\} - E\left\{ \sum_{k=0}^{p-1} w(k)x(n-k) \right\} \cdot x^*(n-k) = r_{dx}k = 0, \ldots, p-1 \]  

And, equation 8 becomes:

\[ \sum_{k=0}^{p-1} w(k)x(n-k) = r_{dx}k = 0, \ldots, p-1 \]  

With \( r_k = \gamma_x(-k) \), the above equation may be written in matrix form:

\[ \begin{bmatrix} r_0 & r_1 & \cdots & r_{p-1} \\
\vdots & \ddots & \ddots & \vdots \\
r_{p-1} & r_{p-2} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} w_0 \\
w_1 \\
\vdots \\
w_{p-1} \end{bmatrix} = \begin{bmatrix} r_{dx0} \\
r_{dx1} \\
\vdots \\
r_{dx(p-1)} \end{bmatrix} \]  

The above matrix equation is actually Wiener-Hopf equation [6], with:

\[ R_xW = r_{dx} \]  

The above Wiener-Hopf equation is finally used for the voice recognition. From equation (13), the input signal \( x(n) \) and the desired signal \( d(n) \) are the only things that we need to know. Then using \( x(n) \) and \( d(n) \), we need to find the cross-correlation \( r_{dx} \). At the same time, using \( x(n) \) finds the auto-correlation \( r_x(n) \) and using \( r_{dx}(n) \) forms the matrix \( R_x \) in Matlab. After computing the \( R_x \) and \( r_{dx} \) we can directly found out the filter coefficients. With the filter coefficients, the minimum mean square-error \( \bar{e} \) can be obtained as:

\[ \bar{e} = E\{e^2(n)\} = E\left\{ \left[ d(n) - \sum_{k=0}^{p-1} w(k)x(n-k) \right]d^*(n) \right\} = r_{dx} - \sum_{k=0}^{p-1} w(k)r_{dx} \]  

In current paper, the first two recorded reference signals can be used as the input signals \( x_1(n) \) and \( x_2(n) \). The third recorded speech signal can be used as the desired signal \( d(n) \). The auto-correlation function of the reference signals is then computed, which are \( r_{x_1}(n) \) and \( r_{x_2}(n) \). Then, the cross-correlation functions for the third recorded voice signal with the first two recorded reference signals are computed, which are \( r_{dx_1}(n) \) and \( r_{dx_2}(n) \). The \( r_{x_1}(n) \) and \( r_{x_2}(n) \) is then used to build the matrix \( R_{x_1} \) and \( R_{x_2} \) respectively. Lastly, using the Wiener-Hopf equation, defined above, the filter coefficients for both two reference signals is calculated and the mean values of minimum mean square-errors with respect to the two filter coefficients is computed. These minimum mean square-errors is then compared and the system give the judgment that the one with smaller mean square error is better recognised signal.

**III. RESULTS & DISCUSSION**

The above mentioned system is implemented in Matlab. The system is tested extensively by different users. For any user, the first two recordings are used as reference signals; whereas, the third recording as target signal. The reference signals corresponds to words “Hello” and “Hi”; whereas the target signal corresponds to word “Hello”. The test is repeated for 10 times for these words to investigate if the judgment of the program is correct. The judgement is expressed in terms of percentage of words recognized correctly. The results corresponding to a male person in the age group 30-35 years are shown below in Figures 2 to 3.
Figure 2 shows the recorded signals corresponding to the words Hello-Hi-Hello. The black and red colored signals are the reference signals and the blue colored signal is the target signal. The corresponding frequency spectrums for three recorded signals are shown in Figure 3. The third signal is repeated for 10 times and corresponding values of errors and frequency offsets are given in Table 1 below.

Table 1: Results corresponding to the three signals Hello-Hi-Hello.

<table>
<thead>
<tr>
<th>Test Times</th>
<th>Frequency_left</th>
<th>Frequency_right</th>
<th>Error1</th>
<th>Error2</th>
<th>Final Recognition Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>2</td>
<td>0.2035</td>
<td>0.2935</td>
<td>“Hello” Recognised</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>10</td>
<td>0.0381</td>
<td>1.128</td>
<td>“Hello” Recognised</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0.2031</td>
<td>0.2732</td>
<td>“Hello” Recognised</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0.2203</td>
<td>0.3112</td>
<td>“Hello” Recognised</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0.5203</td>
<td>1.2203</td>
<td>“Hello” Recognised</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>3</td>
<td>0.293</td>
<td>0.2183</td>
<td>“Hello” Recognised</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>No need</td>
<td>No need</td>
<td>“Hello” Recognised</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0.5203</td>
<td>0.1391</td>
<td>Not Recognised</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0.1173</td>
<td>0.3208</td>
<td>“Hello” Recognised</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0.1103</td>
<td>0.8203</td>
<td>“Hello” Recognised</td>
</tr>
</tbody>
</table>

Total Success Rate =90%

From Table 1, it can be seen that it is difficult to give the judgments with frequency shifts. The frequency shifts are very close between the speech words “Hello” and “Hi”. So the inclusion of FIR Wiener Filter will give better results based on the judgments according to the symmetric errors mentioned in Table 1. When the Error 1 is less than the Error 2, then the words “Hello” are recognised; else, it is not recognised. The results have also shown that when the reference speech signal and the target speech signal are matched, the symmetric errors are smaller. All the judgments made in Table 1 are correct and the overall robustness of the program in terms of success rate is 90%. It has also been noted that judgement based on the frequency shift is not reliable, as the pronunciations of “Hello” and “Hi” are really closed. At this situation, the designed system will give the judgments by comparing the errors of the symmetric property of the cross-correlations. It can also be observed that in some case (for example test series 7 in Table 1), the algorithm consider that the pronunciations of two words are very much different, then the designed system will make the judgments directly by the frequency shifts. The designed system will not calculate the errors in this case. By observing large amounts of simulation results, the author programmed the system in MATLAB to rely frequency shifts when the difference between the absolute values of frequency shifts for the different reference signals is larger or equal to 2. Otherwise, the designed system will continuously calculate the errors of the symmetry.
IV. CONCLUSIONS
From the above made discussions, the designed systems for speech recognition can be easily disturbed by the noise and the way of speaking. The better the signals match, the better is the symmetric property of their cross-correlation, as shown in Table 1. In case of FIR Wiener Filter based speech recognition algorithm, if the reference signal is the same word as the target signal; then using this reference signal to model the target will have less errors. When both reference signals and the target signal are recorded by the same person and in same accent, two systems work well for distinguishing different words, no matter about the identity of the person. But if the reference signals and the target signal are recorded by the different people, both systems don’t work well. So in order to improve the designed systems to make it work better, the further tasks are to enhance the systems’ noise immunity and to find the common characteristics of the speech for the different people.

REFERENCES