National Institute of Technology Karnataka

Digital Processing of Speech and Audio Signals

Speaker Independent Isolated Word Recognotion System

Author:
Anirudh B H (16EC105)
Rithwik Udayagiri (16EC150)
Shivam Kumar (16EC255)

Supervisor:
Dr. Aparna P

November 4, 2018

Contents

Abstract

With the growth of technology and the introduction of smart systems, there is now a need to make the human- machine interface more intuitive and fluidic. One such method is by the introduction of word recognition. This report describes a speaker independent word recognition system that would recognize any given input word that is present in the system's vocabulary. An unknown word is taken as an input. The feature extraction and matching are carried out with the help of Linear Predictive Coding and Dynamic Time Warping which are calculated for multiple time frames. The K Nearest Neighbour Rule is used as the final step for the word recognition In order to improve the accuracy and run-time of the recognition a clustering operation is performed on the dataset.

1 Introduction

Word recognition, both speaker dependent and independent have been extensive fields of research in speech signal processing. The everyday use of word recognition in smart-phones and other electronic gadgets show the success mankind has achieved in this field[1].

Linear Prediction is important in speech analysis because of the accuracy with which it forecasts the time series data. LPC uses linear prediction to model the vocal tract and represent the formants vocal tract structure in a compact manner[1][2]. It is often used for speech compression. The linear predictor coefficients and auto-correlation values can be used to extract formant frequencies, spectral envelope, etc., of the speech signal.

However, the speaker independent recognition of isolated words poses several issues. Each speaker has his own peculiar speech characteristics. They have their own different rates of enunciating a word. They emphasize different parts of the word and words have the influence of their regional accents. These facts lead us to believe that it is impossible to get a high accuracy of recognition when using only one template for a word in the vocabulary. Thus, multiple templates are used for a single word which are obtained from speakers of varying speech characteristics[1]. The above discussion actually points out the key difference between speaker dependent and independent system. Speaker dependent recognition can be achieved with high accuracy using isolated word as a template. Thus, the dichotomy between speaker dependent and independent systems is more of implementation then of structure.

The time taken to utter a word can be different for different people. This leads to different number of time frames of equal length for same words. This can pose a problem during distance calculation between words and templates. Dynamic Time Warping is used to in the recognition process to calculate the distance between unknown word and reference template frame by frame. Frames in perfect registration to each other are found and their distance is added to the overall distance[3].

Accuracy of speaker independent word recognizer is directly proportional to the number of templates being used. Going through each and every template can increase the computation time highly. Thus, sophisticated pattern recognition or clustering algorithms have been used to aid in the optimal selection of the word templates[1][2]. These selected templates represent the variety of ways a single word can be spoken by different speakers.

The purpose of this project is to describe some results on the speaker independent recognition of isolated words based on word templates obtained from a statistical clustering analysis. K nearest neighbour rule is used for deciding the word to be recognized.

2 Theoretical Background

2.1 Pre-Emphasis and LPC Extraction

The word recognizer takes input signal digitized at 16KHz rate, and a p=8^{th} order auto-correlation analysis is performed to calculate LPC coefficients on overlapping frames of N = 720 samples (45 ms), with an overlap of 480 samples (30 ms).

$$H[z] = 1 - az^{-1} \tag{1}$$

where a value of a = 0.96 was used in our simulations[2]. Pre-emphasis serves to reduce the variance of the distance calculations used in the recognition system when LPC parameters are used as the feature set and the auto-correlation analysis is used [citation]. The filtered samples are then sent through a Blackman window. If we denote l^{th} pre-emphasized, windowed frame as x_l(n), 0 ≤ n ≤ N-1, then

$$x_l[n] = x[l * S + n] * w[n] \qquad 0 \leq n \leq N-1, 0 \leq l \leq L-1 \tag{2}$$

where x(n) is the pre-emphasized speech, w(n) is a Blackman window(A hamming window was tested, the Blackman gave a higher result accuracy), S is the shift in samples between adjacent frames, and L is the number of frames in the recording interval.

The next step is to use a p-pole autocorrelation analysis of the word. The value of p=8 was used for this project. These values are used to calculate Linear prediction coefficients which are used for subsequent processing[1].

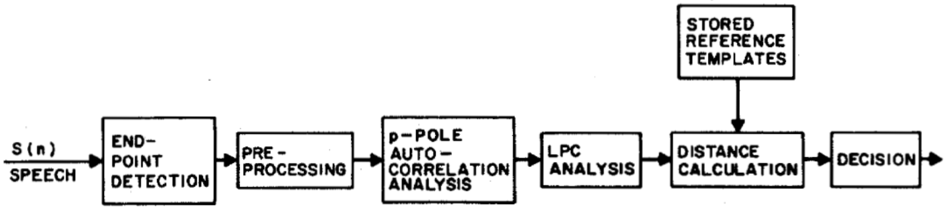

Figure 1: LPC Block Diagram[2]

LPC is a method used to speech processing to extract the features of the speakers. In a pth order LPC system, the current sample is obtained from the previous p samples. If $s[n]$ is the speech sample n and $\hat{s}[n]$ is the predicted sample from the previous p samples, then

$$\hat{s}[k] = a_1 * s[k-1] + a_2 * s[k-2] + \cdots + a_p * s[k-p] \tag{3}$$

where p = 8. In-order to compute the coefficients $a_1, a_2, \ldots a_p$ the first 9 auto-correlation coefficients are calculated for each of the windowed time frames. This is followed by passing the coefficients through the Levison-Durbin Recursion in order to obtain the LPC coefficients.

2.2 Dynamic Time Warping

Dynamic time warping (DTW) is a well-known technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions. Intuitively, the sequences are warped in a nonlinear fashion to match each other. Originally, DTW has been used to compare different speech patterns in automatic speech recognition.

The recognition phase is essentially a matching process in which an unknown sample pattern of,feature coefficients is compared with an ensemble of

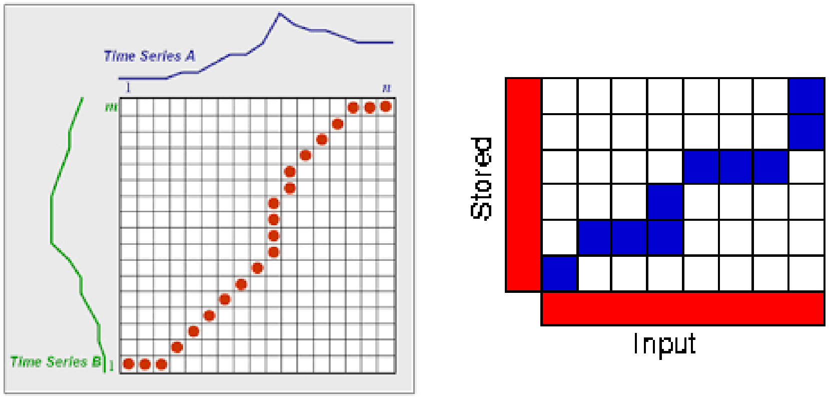

Figure 2: The time alignment between 2 audio files computed by DTW

stored reference templates. The reference templates are from a universal set for speaker-independent and from a particular for speaker-dependent speech recognition. The comparison is carried out on frame by frame scan of the sample pattern against the reference pattern. The distance or dissimilarity measure is calculated using a dynamic programming technique.

The algorithm implemented in simulation is the one proposed by Itakura[3] in which the starting and ending points are assumed to be in perfect registration, and the dynamic path $D(x, y)$ give by :

$$D(i, j) = d(i, j) + min[D(i - 1, j), D(i - 1, j - 1), D(i, j - 1)] \tag{4}$$

where $d(i, j)$ is the distance between the ith frame of the input and the jth frame of the template. The final value $D(n, m)$ gives the overall distance (where n is number of frames in the input and m is the number of frames in the template).

2.3 K Nearest Neighbours

For recognition systems using one or two templates, Nearest Neighbour (NN) rule is used. Here the template whose average accumulated distance D is minimum is chosen as the recognized word. If we denote candidate template words by the index $j, j = 1, 2, \dots, J$, then NN rule is

$$Choose \quad i = i^* \ni D[x, x^{(i^*)}] \leq D[x, x^{(j)}] \qquad 1 \leq j \leq J \tag{5}$$

where $D[x, x^{(j)}]$ is the average distance between the unknown x (input signal) and $x^{(j)}$ (reference template)[1].

A more sophisticated rule is to go for multiple templates. K nearest neighbours (KNN) chooses the vocabulary item as the recognized word whose average distance of K nearest neighbours is minimum. If we denote the sum of K nearest neighbours of j^{th} word to the unknown sample x as $D[x, x_{[k]}^{(j)}]$, then for the KNN rule we compute the quantity r_j as defined in [1]

$$r_j = \frac{1}{K} \sum_{k=1}^{K} D[x, x_{[k]}^{(j)}] \tag{6}$$

j^* is the recognized word, such that

$$r_{j^*} \leq r_j, \qquad j = 1, 2, \ldots, J \tag{7}$$

It should be noted that for K=1, KNN deciding rule becomes NN deciding rule. In this project, we have used K=4 for the K- nearest neighbour deciding rule.

2.4 Clustering

To perform a speaker independent recognition, an appropriate set of features are essential. LPC features are subjective to changes from one speaker to another. Hence there is a need to find a suitable groups amongst each word such that the deviation in a particular group is less while the deviations between different groups is large. Hence a clustering operation is performed.

The clustering here is achieved the K-means clustering[1]. In the first step K random centers are assigned. The distance of each template is calculated with respect the centers are calculated. The template is then assigned to the same cluster as the center with which is has the least distance. The center is re-calculated by taking all the mean of each of the cluster. The process is then repeated till cluster centers become stationary.

The above proposed algorithm would tend to be inefficient if the dataset has outliers or too many deviations. Hence a modified version of the earlier algorithm is implemented. This is the K-median clustering. The algorithm involves the process as the earlier method, with the only difference being the medians are taken as the new centers for the cluster of the members rather than taking the means of the cluster members. This would reduce the effect of outliers and make the algorithm more efficient.

3 Simulation and Results

K mean clustering and K median clustering was calculated for different number of cluster centres, with centres initially randomly assigned. All the obtained LPC coefficients were stored as templates and tested with 1000 test words comprising of words ’forward’, ’backward, ’left and ’right’ spoken by different speakers[4]. Time alignment was achieved using Dynamic Time Warping. DTW distance was found using both MATLAB’s inbuilt dtw function and a custom made function, where MATLAB’s inbuilt dtw had faster run time. Hence, the inbuilt function was used for training purpose of templates using clustering and for final word recognition. Finally, KNN rule was used as the deciding rule to find the unknown audio sample. Different values of K (value used in KNN) were used to find the accuracy of detected word for different number of templates (C) calculated via clustering.

Measure	K = 1	K = 2	K = 5	K = 7
Mean for C = 15	46.48	47.52	47.90	45.33
Median for C = 15	52.10	53.80	57.90	57.24
Mean for C = 25	45.90	49.81	50.19	50.19
Median for C = 25	52.29	57.52	60.00	59.90

Table 1: Accuracy for different values of K and C

The verification and testing was conducted for both mean and median clustering and different value of K and C. The median clustering that was performed was more effective. This is due to the fact that the the dataset that was considered has a certain number of outliers.

It was found that the K = 5 and C = 25 yields a higher accuracy compared to the other. There is an increase in the accuracy by 4 to 8 percentage between K = 1 and K = 5. A larger value of K would result in the consideration of the outliers for the minimum distance calculations.Hence the value is K is chosen neither too high nor too low. As we increase the cluster till 25 the accuracy increases, beyond which the accuracy either is constant or decreases. An increase in C from 15 to 25 results in a increase in accuracy. This would lead to the result that the clustering helps improve the accuracy

The LPC Coefficients are by nature dependent on the speaker. Hence even with the use of basic clustering algorithms use of them in speaker independent word recognition would be less effective and would yield a moderate accuracy. The accuracy however can be improved by superior clustering and speech enhancement algorithms.

4 Conclusion and Future Work

In this paper a speaker independent word recognition algorithm is proposed. The proposed algorithm makes used of the LPC features to find the similarities between two spoken words. However LPC features are subjective to each speaker. This would cause a reduction in the accuracy. In-order to increase the accuracy the input would needed to be compared with multiple templates. This however would be expensive for large data-sets. Hee a clustering is performed to achieve better results. The highest accuracy achieved is 60% .

Various other methods such as constrained DTW, end-point detection and other superior clustering algorithms can be used to improve the accuracy. Multiple candidates for the clustering can be taken form each cluster. This would increase the accuracy of the overall system.

References

[1] Rabnier et.al, "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques", IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOE. ASSP-27, N o . 4,AUGUST 1979

[2] V. N. Gupta, J, K. Bryan, and J. N.Gowdy, "A speaker-independent speech recognition system based on linear prediction" IEEE Truns. A must., Speech, Signal Processing, vol. ASSP-26, pp. 27-33, Feb. 1978.

[3] F. Itakura, "Minimum prediction residual applied to speech recognition" IEEE Trans. Acoust., Speech, Signal Processing, vof. ASSP-23, pp. 67-72, Feb. 1975.

[4] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," arXiv preprint arXiv:1804.03209, 2018.