International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 4 840 - 843

Real Time Home Automation Using Speech Recognition

Abhay K Khalane
Department of ECE engineering
LNCT college of Indore [MP]
abhay.khalane@gmail.com

Asst. Prof. Neha Namdev
Department of ECE engineering
LNCT college of Indore [MP]
softneha7@gmail.com

Prof.A C Tiwari
Head of Department of ECE engineering
LNCT college of Indore [MP]
achandra0@gmail.com

Abstract: This paper concludes about a security system that is implemented over the technological concepts of speaker identification. Mel frequency cepstral Coefficients {MFCCs} are the key parts used for feature extraction. And vector quantization techniques are integrated to lower the amount of data to be handled. In speech recognizers MEL's can be used as to parameterize the speech. Practically, when implemented speech recognition and dialogue systems sometimes might introduce a need to synthesize and or reconstruct the speech from the MFCCs which has already been transmitted or already saved.

Keywords: *MFCC (Mel-frequency Cepstrum Coefficient), Feature Extraction, Feature Matching, dynamic Time Warping (DTW).*

*****

I. INTRODUCTION

Speech can be considered as the most common known example of natural communication. Recently the developments have been made to use this in the security system and it has been proved much more effective as well. Voice Signal Identification includes the processes to convert a speech waveform into samples that are useful for further processing using the available algorithms and there are many of them. The first step includes conversion of human voice into digital signal form. This is done at every discrete step to obtain samples at various different time steps. The digitized speech samples then can be processed using various algorithms like MFCC. After which, the coefficient of voice samples are passed through DTW to choose the pattern that is the perfect match to that with the database and input frame in order to minimize the conflict which might happen between them. The most widely used cepstrum based methods to compare the patterns and the similarities are the MFCC and DTW. The MFCC and DTW techniques can be implemented using MATLAB. This paper explains the inventions of the voice recognition study using the MFCC and DTW techniques.

II. PRINCIPLE OF SPEECH RECOGNITION

Speech Recognition Algorithms

A voice analysis can be done after taking an input of a user through microphone. The design of the system is such that manipulation of the input audio signal can be done. At different levels, different operations can be performed on the input signal such as Pre-emphasis, Windowing, Framing, Mel Cepstrum analysis and Recognition (Matching) of the spoken word is also done. The voice algorithms include two distinguished and separate phases. The first one is training sessions, and the second one is said to be as operation session or testing phase as described in figure 1.

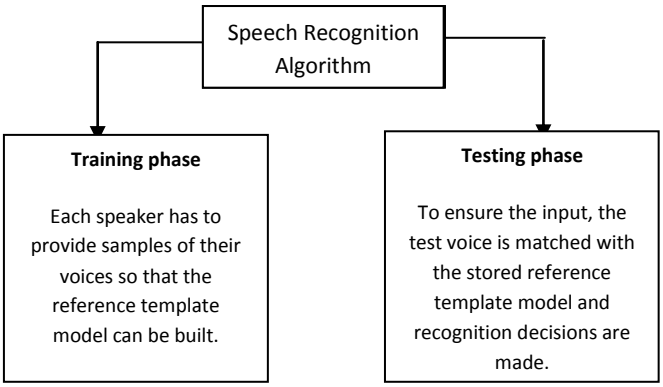


Fig.1. Speech Recognition algorithms

2.1 Feature Extraction (MFCC)

The extraction of the parametric representation of acoustic signals is important task to be achieved so as to produce a clear and better recognition performance. The overall correctness and efficiency of this phase is important to consider for the next phase since it effectively affects its behavior. MFCC is based on human hearing capabilities which apparently cannot read the frequencies over 1Khz. In other words, MFCC based on known variation of the human ear's critical bandwidth reading capability with frequency [8-10]. MFCC has two types of filters which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on the Mel Frequency Scale to capture important characteristic of phonetics in speech. The overall process of the MFCC is shown in Figure 2.

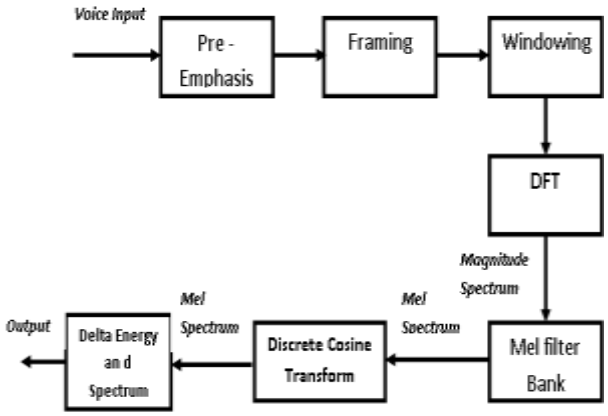


Fig.2. MFCC Block Diagram

As shown in Figure 3, MFCC consists of the coefficient computational step. Each step has its function and the mathematical approaches as discussed briefly in the following:

Step 1: Pre–emphasis

This step processes the passing of signal through a filter which emphasizes highest frequencies. This process will be the increase the energy of the signal at higher frequency.

Lets consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

$$Y[n] = X[n] - 0.95\,X[n-1]$$

Step 2: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 21 to 40 msec. The voice signal is divided into frames of N samples. Adjacent the frames are being separated by M (M<N). Typical values used are M = 100 and N= 256 named Times. Right margins should be justified not ragged.

Step 3: Hamming windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window is equation given as: If the window is defined as W (n), $0 \le n \le N-1$ where N = number of samples in each frame

Y[n] = Output signal

X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y(n) = X(n) \times W(n)$$

Step 4: Fast Fourier Transform

Fast Fourier Transform to convert each frame of N samples from time domain into the this frequency domain. Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the frequency time domain. This statement supports the equation below:

$$Y(w) = FFT\ [h(t) * X(t)] = H(w) * X(w)$$

Step 5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear 0scale. The bank of filters according to Mel scale as shown in figure 4 is then performed.

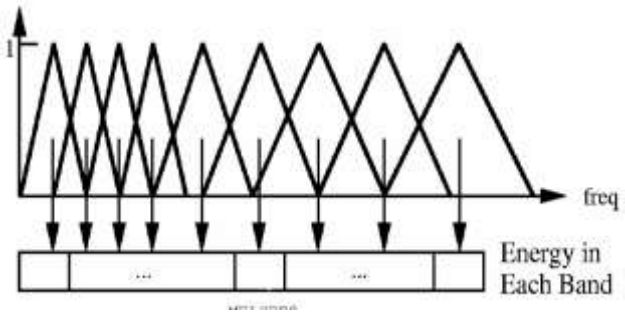


Fig. 3.Mel scale filter bank, from (young et al1997)

This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude the frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [7, 8]. Then, each filter output is the sum of its filtered spectral components. After that following equation is used to compute the Mel for given frequency f in HZ:

Step 6: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of this coefficient is called acoustic vectors. Therefore each this input utterance is transformed into a sequence of acoustic vector.

Step 7: Delta Energy and Delta Spectrum

The voice signal and the frames changes, such as the slope of the formant at its transitions. Therefore, this is a need to add features related to the change in cepstral features over time .13 delta or velocity features (12 cepstral features plus the energy), and 39 features a double delta or the acceleration feature are added. The energy in a frame for

the signal x in a window from time sample t1 to time sample t2, represented at the equation below:

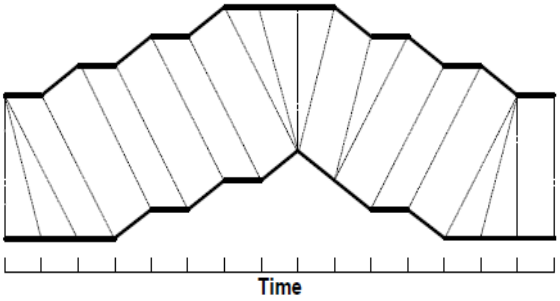
$$Energy = \sum X^2[t]$$

Each of the 13 delta feature represents the change between frames in the equation 8 corresponding cepstral or energy feature, while each of 39 double delta is features represents the change between frames in the corresponding delta features.

$$d(t) = \frac{c(t+1) - c(t-1)}{2}$$

Feature Matching (DTW):

DTW algorithm is based on Dynamic Programming techniques as describes in [11]. This algorithm is for measuring similarity between two time series which may vary in time or speed. This technique also used to find optimal alignment between the two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. This warping between the two time series it can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series. Figure 4 shows the example of how one times series is 'warped' to another.



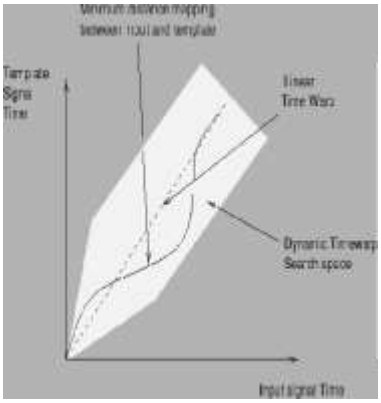


Fig. 4. A Warping between two time series

In Figure 4, every vertical line interfaces a point in one time arrangement to its correspondingly comparative point in the other time arrangement. The lines have comparative this quality on the y-hub, however have been isolated so the vertical lines between them can be seen all the more effortlessly. In the event that both of the time arrangement in figure 4 was indistinguishable, every one of this line would be straight vertical lines on the grounds that no twisting would be important to 'line up' the two time arrangement. The twist way separation is a measure of the contrast between the two time arrangement after they have been distorted together, which is measured by the entirety of the separations between every pair of focuses associated by the vertical lines in Figure 4. Along these lines, two time arrangement that are indistinguishable with the exception of restricted extending of the time pivot will have DTW separations of zero. The guideline of DTW is look at two element examples and measure its similitude by ascertaining a base separation between them. This great DTW is processed as beneath.

$$Q = q1, q2, \ldots, qi, \ldots, qn$$
$$C = c1, c2, \ldots, cj, \ldots, cm$$

To align two sequences using DTW, an *n*-by-*m* matrix where the (*i*th, jth) element of the matrix contains the distance *d* (*q*i, *c*j) between the two points *q*i and *c*j constructed. Then, absolute distance between the values of two sequences is calculated using the Euclidean distance computation.

Each matrix element (*i*, *j*) corresponds to the alignment between the points *q*i and *q*j. Then, accumulated distance is measured by:

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

This is appeared in Figure 5 where the even pivot speaks to the season of test information signal, and the vertical hub speaks to the time succession of the reference format. The way demonstrated results in the base separation between the information and layout signal. The shaded in range speaks to the quest space for the information time to layout time mapping capacity. Any monotonically non-diminishing way inside the space is another option to be considered. Utilizing the dynamic programming systems, the quest for least separation way should be possible in the polynomial its time P (t), utilizing condition beneath:

$$P(t) = O\left(N^2 V\right)$$

Fig.5. Example Dynamic time warping (DTW)

Where, N is the length of sequence, and V is the number of templates to be considered.

Theoretically, the major optimizations to the DTW algorithm arise from observations on the nature of good paths through
the grid. These are outlined in Sakoe and Chiba [16] and can be summarized as:

***Monotonic condition*:** the path will be not turn back on itself, both I and j indexes either stay the same or increase, they never decrease.

***Continuity condition*:** The path advances one step at a time. Both i and j can be only increase by 1 on each step along the path.

***Boundary condition*:** the path start at the bottom left and ends at the top right.

***Adjustment window condition*:** a good path is unlikely to wander very far from the diagonal. The distance that the path allowed to the wander is the window length *r*.

***Slope constraint condition*:** The path should not be too steep or too shallow. This prevents very short sequences the matching very long ones. The condition is expressed as the ratio n/m where m is the number of step in the x direction and m is the number in the y direction. After m step in x you must make a step in y and vice versa.

CONCLUSION

This paper focuses on two voice recognition algorithms techniques which are important in improving the voice recognition performance. The algorithm techniques are able to authenticate the particular speaker based on the individual information that is included in the voice signal. These techniques could be used effectively for voice recognition purposes. Several other techniques such as Liner Predictive Coding (LPC), Hidden Markov Model (HMM), and artificial Neural Network (ANN) are currently being investigated. The findings will be the presented in future publications.

ACKNOWLEDGMENTS

The authors would like to the thanks Department of ECE engineering for supporting this survey.

REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. .

[2] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[3] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

[4] Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[8] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[9] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender