DCT APPLICATION IN SPEECH RECOGNITION: A SURVEY

Atul Narkhede[1], Dr. Naveen Sen[2], Dr. Milind Nemade[3]
1(Research Scholar, Faculty of Engineering, Pacific Academy of Higher Education and Research University, Udaipur atulnarkhede22@gmail.com)
2(Associate Prof., Pacific Academy of Higher Education and Research University, Udaipur naveensen1984@gmail.com)
3(Professor, Department of Electronics Engineering, K. J. Somaiya Institute of Engineering & Information Technology, Mumbai, India. mnemade@somaiya.edu)

Abstract: - Speech recognition with the help of the machine is automatically an important research area for over forty years. Since the voice is an unlimited information signal, the speech signal processing through digital conversion is a very efficient tool for high and accurate automatic signal or voice recognition technology. Speech recognition has found its application in different areas of our daily life as a telephone answering machine for transmitting text and sending voice signals to machines. Function extraction and classification is a major part of the ASR system process. The main part of the voice processing system to improve capacity is the selection of the function extraction method that plays an important role in the accuracy of the system. This document provides a brief overview of the detection of various methods in speech processing where DCT uses to efficiently extract features in different ways.

Keywords: DCT, MMSE, MFCC.

I. Introduction

Automatic speech recognition by machine has been a research goal for over four decades. In the world of science, the computer has always understood human mimics. The idea that was generated to make the speech recognition system is because it is convenient for humans to interact with a computer, a robot or any machine by voice or vocalization instead of difficult instructions. Humans have long been inspired to create a computer that can understand and speak like humans. Speech recognition is the process by which the computer assigns an acoustic voice signal to some form of abstract vocal meaning. This process is very difficult, since the sound must correspond to the fragments of sound stored in which a subsequent analysis must be performed because the fragments of sound do not correspond to the pre-existing sound pieces. Various methods of feature extraction and model matching techniques are used to create better quality speech recognition systems. The feature extraction technique and model matching techniques play an important role in the voice recognition system to maximize the speech recognition rate of different people. Following are some methods that explain the advantages and disadvantages of DCT.

II. DCT for noise Reduction:

This article illustrates the advantages of using the discrete cosine transform (DCT) over the discrete standard Fourier transform (DFT) in order to eliminate the noise embedded in a voice signal. The derivation of the minimum mean square error filter (MMSE) based on the statistical modeling of the DCT coefficients is shown. The derivation of an excessive attenuation factor is also demonstrated by the fact that the speech energy is not always present in the noisy signal at all times or in all coefficients. This excessive attenuation factor is useful for suppressing any residual musical noise that may be present. It is often necessary to improve speech by eliminating noise in voice processing systems operating in noisy environments. The energy of

white noise is uniformly distributed throughout the spectrum, but the energy of speech, particularly of sound, is concentrated in certain frequencies. Therefore, the advantage of using a real transformation, like the DCT considered in this document, is that the problem of not correcting the phase will have less serious consequences. DCT is widely used in image compression due to its excellent energy compaction property. This is also a useful function to eliminate noise. DCT provides significantly higher energy compaction than DFT [1].

III. Hybrid Method Genetic-Fuzzy Inference System:

In this system, a voice signal is coded and parameterized in a two-dimensional time matrix, with four parameters of the voice signal. After encoding, the mean and variance of each model is used to generate the rule base of the fuzzy inference system Mamdani. The mean and variance are optimized using a genetic algorithm to obtain the best performance of the recognition system. Consider the Brazilian expressions (digits) as schemes: 0,1, 2,3,4,5, 6,7,8,9. The discrete cosine transformation (DCT) is used to encode vocal patterns. The use of DCT in data compression and model classification has increased in recent years, mainly because its performance is much closer to the results obtained from the Karhunen-Lo` transformation, which is considered optimal for a variety of criteria such as mean square error of truncation and entropy. This article demonstrates the potential of DCT and the fuzzy inference system in speech recognition. These two tools have shown good results in the temporal modeling of the vocal signal [2].

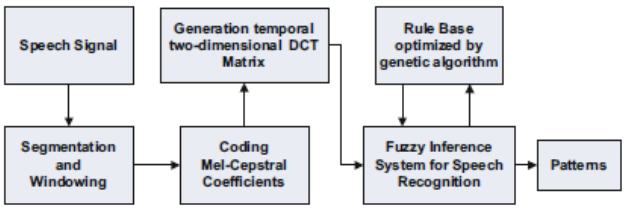

Fig. 1. Block diagram of the proposed recongnition system HMFE

IV. MMSE filter using DCT:

This method illustrates the properties of the discrete cosine transformation (DCT) with respect to the discrete standard Fourier transformation (DFT) in the case of elimination of speech noise. The results show that DCT has better energy compaction and fewer calculations than DFT. The proposed algorithm is implemented for the reduction of residual noise using the probability of the absence of speech technique. The proposed techniques use adaptive schemes that will monitor the probability of the absence of speech in a noisy speech. Estimate the spectral width received from a binary classification, that is, speech is present or absent.

The presence of different noises such as:

- Background noise
- Channel noise
- Quantization noise

It significantly degrades system performance, such as voice encoders and speech recognition systems, so we have to do a preprocessing step in these systems that incorporate speech enhancement to eliminate noise. The filtering process must be performed to filter a signal and eliminate noise. So we can define the processing of the filter as follows: The information extraction process that carries the X (n) signal from the observed signal Y (n), where Y (n) = X (n) + N (n) and N (n) is a noise process, it is called a filter. Different algorithms are used both in the time and frequency domain to eliminate the noise embedded in the noisy voice signal [3].

V. DCT and MFCC:

This paper examines and presents an approach for speech signal recognition using frequency spectral information with Mel frequency. It is a dominant feature for speech recognition. The mel coefficients of Cepstral (MFCC) are the coefficients that collectively represent the short-term power spectrum of a sound, based on a linear transformation of the cosine of a logarithmic power spectrum on a non-linear mel frequency scale. The performance of the MFCC is influenced by the number of filters, the shape of the filters, the filter spacing mode and the deformation of the power spectrum. In this document, the optimal values of

the above parameters are chosen to obtain an efficiency of 99.5% in a very small audio file length.

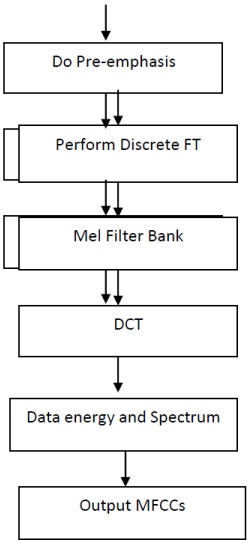

Figure 2. Process model for extracting MFCCs from an audio speech

a) Pre-emphasis: normally, a FIR filter of a coefficient is known as a pre-emphasis filter.
b) Framing: frames generally have 20-30 ms with an overlap of 10-15 ms.
c) Windows: you can use the functions of the Hamming or Hanning window.
d) DFT: to convert each frame of N time domain samples into the frequency domain.
e) Mel filtering: the magnitude frequency response of each filter has a triangular shape and is equal to the unit at the central frequency and decreases linearly to zero at the central frequency of two adjacent filters.
f) DCT: this is the process to convert the spectrum of Mel records into the time domain using DCT. The result of the conversion is called the Mel coefficient of Cepstrum. The set of coefficients is called acoustic vectors. Therefore, each incoming emission is transformed into an acoustic vector sequence [4].

In this survey, they focus on providing better performance in the speech recognition algorithm by integrating digital signal transposition with voice recognition techniques. This is an approach to improve the performance of the speech recognition algorithm by using the Butterworth stop band filter and the voice compression based on the discrete transformation of the cosine with inverse wave transformation. The main objective is to integrate the filter with the voice recognition algorithm to improve the results when there is noise in the signal. In this work, the correspondence is made using inverse wave transformations that reduce the speech recognition time. The proposed algorithm is designed and implemented in MATLAB. The proposed algorithm was tested on the samples provided and evaluated using different recognizable and unrecognizable samples, obtaining a recognition ratio of approximately 98%. It has been shown that the proposed algorithm provides better results than existing techniques. The proposed algorithm increases the accuracy of the voice recognition system. In the proposed method, the goal is to detect the speaker from previously recorded wave samples. The main concentration is in precision and speed. The proposed method is implemented using MatLab.

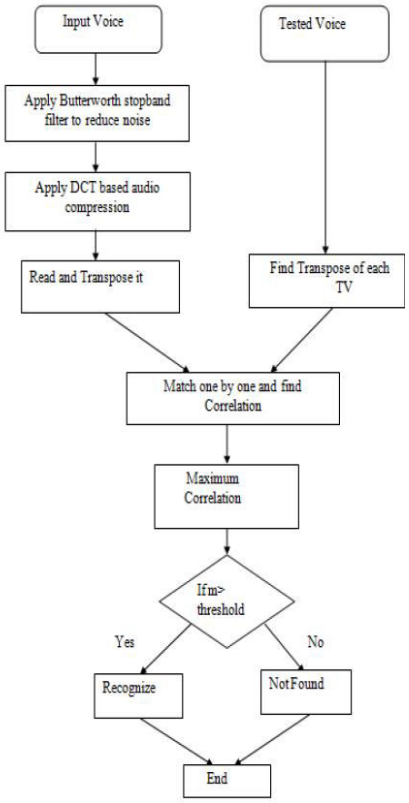

Fig.3. Flow-Chart For Speech Recognition Algorithm

Voice compression based on discrete cosine transformations (DCT) is used to reduce the size of vocal information. It is used to speed up the system by eliminating the redundancy of audio information. Compression is the process of eliminating redundancy and duplicity. DCT is very common when encoding video and voice tracks on computers [5].

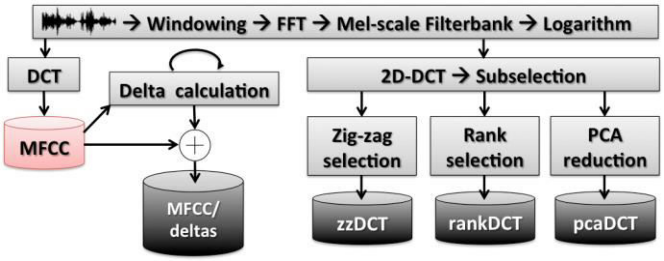

Fig. 4. Feature extraction methods

VI. 2D DCT:

This proposed method used the coefficients extracted from the discrete cosine transform 2D (DCT) of the energies of the Log Mel filter bank to improve the recognition of the diffusers on the traditional Mel cepstral frequency coefficients (MFCC) with delta and double deltas (MFCC / delta). The selection of the relevant coefficients proved to be crucial, which led to the proposal of a zigzag analysis strategy. Although the 2D-DCT coefficients have provided significant gains on MFCC / delta, the analysis strategy remains sensitive to the number of outputs of the filter bank and to the size of the analysis window. In this work, we analyze this sensitivity and propose two new data-based methods to use the DCT coefficients for the recognition of the speakers: rankDCT and pcaDCT. The first, rankDCT, is an automatic coefficient selection strategy based on the highest average intra-frame energy range. The alternative method, pcaDCT, avoids the need for selection and instead projects the DCT coefficients on the desired dimensionality through the analysis of the main components (PCA). All functions, including MFCC / delta, are set in a subset of the PRISM database to subsequently highlight the sensitivity of the parameters of each function. Evaluated in the recent NIST SRE'12 corpus, pcaDCT constantly exceeds the characteristics of the DCT and zzDCT range and offers an average relative improvement of 20% on MFCC / delta in all conditions [6].

Another approach is a 2D DCT-based approach to compress the acoustic characteristics for remote speech recognition applications. The coding scheme involves the calculation of a 2D DCT in blocks of feature vectors followed by uniform scalar quantification, stroke length and Huffman coding. Digit recognition experiments were conducted in which the training was conducted with un-quantified cephalic features of clean voice and the tests used the same characteristics after encoding and decoding with 2D DCT and entropy coding and at various noise levels. Acoustic the coding scheme translates into recognition performances comparable to those obtained with characteristics that are not quantified at low bit rates. MFCC's 2D DCT coding together with a method for analyzing variable frame rates [Zhu and Alwan, 2000] and peak isolation [Strope and Alwan, 1997] maintains the noise robustness of these low SNR algorithms even at 624 bps

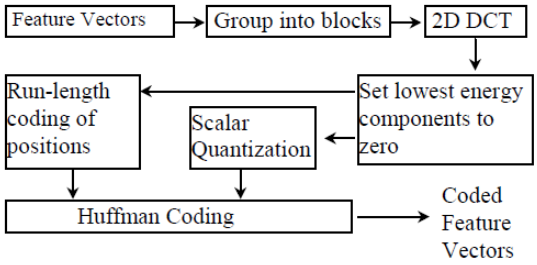

Figure 5. Block diagram of the DCT and entropy encoder.

In the client, the entry is first segmented into frames, the characteristics are calculated for each frame and then feature blocks are generated. A 2D DCT is then performed on each block and the components with the lowest energy are set to zero. This is followed by scalar quantification, execution length and Huffman coding. A block diagram of the encoder is shown in Figure 5. In the receiver

decoding and IDCT are performed and in the ASR system the characteristic vectors corresponding to each frame are inserted. Only function vectors are encoded and sent to the recognition server; the first and second derivatives are calculated on the server based on the features retrieved [7].

VII. BDCT Method:

Robust speech recognition has become an important area of research in recent years. Multi-band functions can be combined in different ways to perform the speech recognition task. The extraction of multiband characteristics will propose a transformation of the cosine to discrete blocks (BDCT) with its transformation matrix of the nucleus derived from the decomposition of the discrete cosine transformation nucleus (DCT). We show that the BDCT approaches the DCT to maintain information in the correlation of a sequence. When the BDCT is applied to the energies of the mel filter bank frequency (FBE) to replace the DCT to convert them into cephalic coefficients, a new type of MFCC is produced [8].

VIII. Conclusion

This Paper briefly explain different methods used for speech recognition using DCT, which shows that DCT can be used for noise reduction very well, also it has property of energy compaction which can improve speed as well as recognition rate.

References:

[1] Ing Yann Soon*, Soo Ngee Koh, Chai Kiat Yeo, "Noisy speech enhancement using discrete cosine transform ," ELSEVIER, Speech Communication, pp 249 – 257,1998

[2] Washington Silva and Ginalber Serra, " An Intelligent System Based on Discrete Cosine Transform for Speech Recognition," ResearchGate, IBERAMIA, LNAI 7637, pp. 320–329, November 2012.

[3] Muhammad Safder Shafi, Mansoor Khan, "Transform Based Speech Enhancement Using DCT Based MMSE Filter,& Its Comparison With DFT Filter," Journal of Space Technology, Vol 1, No. 1, pp 47 – 52, July 2012.

[4] Garima Vyas, Barkha Kumari, "Speaker Recognition System Based On MFCC and DCT," International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-5, pp 167 – 169, June 2013.

[5] Sukhdeep Kaur, Er. Gurwinder Kaur, "Enhancement of Speech Recognition Algorithm Using DCT and Inverse Wave Transformation," International Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 3, Issue 6, pp.749-754, Nov-Dec 2013.

[6] Mitchell McLaren, Yun Lei, "Improved Speaker Recognition Using DCT Coefficients as Features,"IEEE International Conference on Acoustics, Speech and Signal Processing, 978-1-4673-6997-8, pp 4430 – 4434, April 2015.

[7] Qifeng Zhu and Abeer Alwan, "An Efficient And Scalable 2d Dct-Based Feature Coding Scheme For Remote Speech Recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, ISSN 1520 – 6149, May 2011.

[8] Suman K. Saksamudre, R. R. Deshmukh, "Comparative Study of Isolated Word Recognition System for Hindi Language," International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 4 Issue 07, July-2015.