RESEARCHERID
THOMSON REUTERS

[Jolad * *et al.*, 7(2): February, 2018]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

DIFFERENT FEATURE EXTRACTION TECHNIQUES FOR AUTOMATIC SPEECH RECOGNITION: A REVIEW

Bhuvaneshwari Jolad[*1] & Dr. Rajashri Khanai[2]

[*1]Research Scholar, KLE Dr. MSSCET, Belagavi, India

[2]Department of Electronics and Communication Engineering, KLE Dr. MSSCET, Belagavi, India

DOI: 10.5281/zenodo.1165872

ABSTRACT

Automatic speech recognition, which allows a usual and user-friendly communication technique among individual and device, is a dynamic research area. The speech recognition is the skill to pay attention to what we are talking about, to interpret and to perform actions based on the information spoken. This article presents a short outline of speech recognition and the various techniques like MFCC, LPC and PLP intended for feature extraction in speech recognition system. Among the three techniques i.e. MFCC, LPC, PLP, Mel frequency cepstral coefficient's (MFCC) is repeatedly used feature extraction technique in speech recognition process because it is most nearby to the real individual acoustic speech opinion.

KEYWORDS: Automatic Speech Recognition, Feature Extraction, MFCC, LPC, PLP.

I. INTRODUCTION

Speech is the most common type of individual communication and one of the most exciting investigation areas of the signal processing is speech processing. Speech processing is nothing but learning of language signals and the processing techniques of these signals. The signals are usually processed in a digital version, so speech processing can be viewed as a unique case of digital signal processing which is applied to speech signal.

Speech Recognition is one of the plunge investigation areas in language (speech) processing, which is also known as automatic speech recognition (ASR). Speech recognition technology allows a computer to pay attention to individual voice commands and to understand individual languages. Speech recognition is the procedure of altering a given input signal into a series of words by means of an algorithm that is implemented as a computer program. That is, the speech recognition system enables a computer to recognize the words an individual speaks in a microphone or phone and convert it into readable text. Speech Recognition has numerous applications such as in health care, military, helicopters, telephony and other domains etc.

Advancement in language (speech) technology was motivated as people wanted to develop mechanical models that allow the emulation of individual oral announcement abilities. Computers use speech processing to track voice commands and diverse individual languages.

Figure 1 shows a fundamental form of speech recognition system that denote diverse phases of a scheme that contains pre-processing, speech feature extraction, classification and language model [1].

The input signal will be changed by preprocessing stage before any information can be extracted at feature extraction phase. The feature extraction phase extracts essential vectors needed for use in modeling phase after preprocessing phase. The extracted vectors have to be strong (robust) to noise for improved accuracy.

The language text is recognized by classification phase by using extracted features and a language template where the language template contains syntax and semantics related to the responsible language which help the classifiers to identify the input statement [1].

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

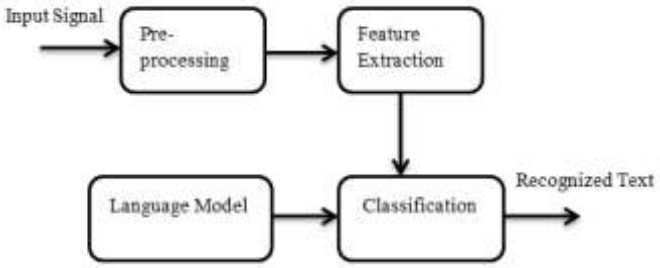

Fig.1 Stages of Speech Recognition

II. FEATURE EXTRACTION TECHNIQUES

Feature extraction acts as a exceptionally important job in sspeech recognition procedure and as it draws out valuable data from sample speech it is a vital part of research for many years. The key objective of this method is to find out the performance level of different feature extraction techniques and then selecting one of the methods among them. It plays an important role in accuracy of speech recognition.

To separate one speech signal from the other feature extraction technique plays an important role. Since each language has different individual characteristics implanted in statement, these characteristics can be extracted from an extensive variety of feature extraction techniques proposed and effectively exploited for speech recognition task. Some significant feature extraction techniques for speech recognition system are explained in next section:

Mel Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) is most widespread and important technique used to extract spectral features. MFCCs used in speech recognition are based on frequency domain using the Mel scale which is based on the human ear scale and they are one of the most accepted feature extraction techniques. MFCCs which are well thought-out to be frequency domain features are to a great extent more precise than time domain features [5], [6].

Human Speech as a function of the frequencies is not linear in nature; therefore the pitch of an acoustic speech signal of single frequency is mapped into a "Mel" scale. In Mel scale, the frequencies spacing below 1 kHz is linear and the frequencies spacing above 1 kHz is logarithmic [27]. The Mel frequencies corresponding to the Hertz frequencies are calculated by using equation (1).

$$fmel = 2595 * \log(1 + \frac{f}{700}) \quad (1)$$

The block diagram for Mel-Frequency Cepstral Coefficients (MFCC) computations is shown in Fig. 2.

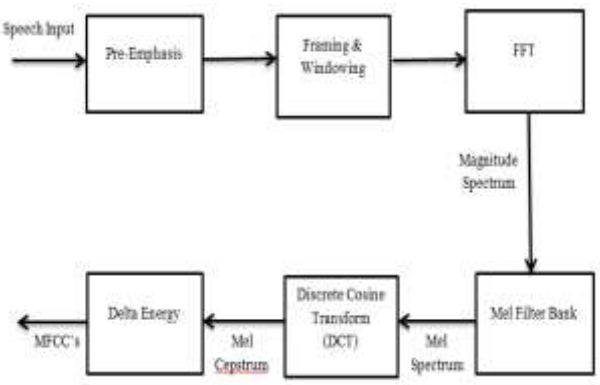

Fig. 2 Block Diagram for MFCC Computation

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

The inner blocks shown in Fig. 2 are individually described below in detail:

1) Pre-Emphasis: The audio signals are recorded having a sampling rate of 16 kHz. Each word is stored in separate audio file. This step includes the Pre-emphasis of signal to boost the energy of signal at high frequencies. The difference equation of Pre-emphasis filter is given by equation (2).

$$H(z)=\frac{B(z)}{A(z)}=\frac{\left(b_0+b_1z^{-1}\right)}{1}=1-0.97z^{-1} \qquad (2)$$

2) Framing and Windowing: The language (speech) signal is not stationary in nature. In order to make it stationary framing is used. Framing is the next step after pre-emphasis; in this step speech signal is split up into smaller frames overlapped with each other. After framing, windowing is used to remove discontinuities at edges of frames. Hamming Window is the window method used in this research .The Hamming Window is given by the equation (3).

$$w(n)=\begin{cases}0.54\text{-}0.46\cos\left[\frac{2\pi n}{N\text{-}1}\right] & 0\le n\le N-1\\ 0 & otherwise\end{cases} \qquad (3)$$

Where, N is total quantity of samples in a single frame.

3) Fast Fourier Transform (FFT): Fast Fourier transform is used for calculating of the discrete fourier transform (DFT) of signal [6]. This step is performed to transform the signal into frequency domain. The FFT is calculated using equation (4).

$$x[k]=\sum_{n=0}^{N-1}x(n)e^{-j2\frac{\pi}{N}kn} \qquad (4)$$

Where, N is the size of FFT.

4) Mel Filter Bank: The next step is transformation from Hertz to Mel Scale, the spectrums power is transformed into a Mel scale [7]. The Mel filter bank consist of triangular shaped overlapping filters.

5) Discrete Cosine Transform (DCT): The Discrete Cosine Transform (DCT) is employed after taking logarithm of output of the Mel-filter bank.

6) Delta Energy: In this step take base 10 Logarithm of output of previous step. The computation of Log energy is essential because of the fact that human ear response to acoustic speech signal level is not linear, human ear is not much sensitive to difference in amplitude at higher amplitudes. The advantage of logarithmic function is that it tends to duplicate behavior of human ear. Energy computation is calculated using equation (5)

$$E=\sum_{t=t1}^{t=t2}x^2(t) \qquad (5)$$

It finally produces the Mel frequency cepstral coefficients.

Linear Predictive Coding (LPC)

Linear predictive coding is an arithmetical computational process which is linear mixture of numerous preceding samples. Linear predictive coding [7] [8] of language (speech) is the major method for approximating the fundamental parameters of language. It gives a precise approximation of language parameters and is also a

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

competent computational model of language. The fundamental thought behind LPC is that a language sample can be estimated as a linear mixture of past language samples. A single set of parameters or predictor coefficients can be determined by decreasing the addition of squared variances amid the real language samples and expected values. The coefficients so obtained become the basis for linear predictive coding of language [10].

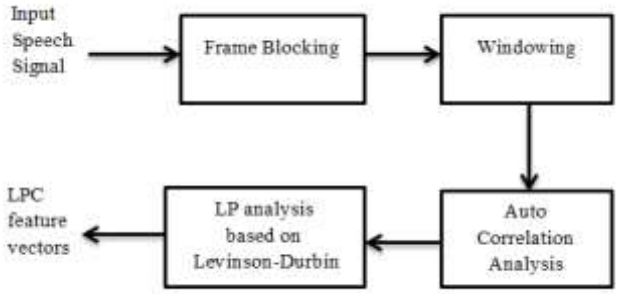

Fig.3 Block diagram of Linear Predictive Coding

The block diagram for linear predictive coding is shown in Fig.3. The speech features are extracted by processing the speech signal that is sampled directly from microphone. The technique meant for feature extraction procedure is Linear Predictive Coding using Linear Predictive Coding (LPC) Processor. Following are the fundamental steps of Linear Predictive Coding (LPC) processor [31, 32]:

1. Pre-Emphasis: Here the digitized language (speech) signal, s(n), is sent through a low order digital system, to spectrally flatten the signal and to make it less prone to finite precision effects later in the signal processing. The pre-emphasizer network output is related to the input of the network, s(n), by the following equation:

$$\tilde{s}(n) = s(n) + \alpha \tilde{s}(n-1) \qquad (1)$$

2. Frame Blocking: The output of previous step is blocked into frames of N samples, where the neighboring frames are separated by M samples. Let us assume that, the l^{th} frame of language be $x_l(n)$, and let there be L frames inside entire language signal, then $x_1(n)$ is given as

$$\tilde{x}_l(n) = \tilde{s}(Ml + n) \qquad (2)$$

In which, n = 0, 1, …, N – 1 and l = 0, 1, …, L – 1

3. Windowing: Here in this step every individual frame is windowed to reduce the signal discontinuities at the start and end of each frame. If we define the window as w(n), in which $0 \le n \le N - 1$, then the consequence of windowing is given by the signal:

$$\tilde{x}_l(n) = x_l(n)w(n) \qquad (3)$$

In which, $0 \le n \le N - 1$

Hamming window is the typical window that takes the form

$$w(n)=0.54-0.46\cos\left[\frac{2\pi n}{N-1}\right] \qquad (4)$$

In which, $0 \le n \le N - 1$

4. Autocorrelation Analysis: To auto correlate every frame of windowed signal is the next step, in order to give

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m) \qquad (5)$$

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

In which, m = 0, 1, …, p

5. Linear predictive coding (LPC) analysis: It is the next processing step that translates every frame of p+1 autocorrelations into linear predictive coding (LPC) parameter set by using Durbin's method. This is officially given as the following algorithm:

$$E^{(0)} = r(0) \tag{6}$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} r(|i-j|)}{E^{i-1}} \qquad 1 \le i \le p \tag{7}$$

$$\alpha_i^{(i)} = k_i \tag{8}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \tag{9}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{i-1} \tag{10}$$

Equations (6) to (10) are solved recursively for i = 1, 2, …, p, to get the Linear predictive coding (LPC) coefficient, a_m , as

$$a_m = \alpha_m^{(p)} \tag{11}$$

6. Linear predictive coding (LPC) parameter conversion to cepstral coefficients: A very important LPC parameter set is LPC cepstral coefficients that can be derived straight from the LPC coefficient set. Following is the recursion used

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k} \qquad 1 \le m \le p \tag{12}$$

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k} \qquad m > p \tag{13}$$

Features that are extracted from speech signal are the LPC cepstral coefficients and these are fed as input data for speech recognition system.

Perceptual Linear Prediction (PLP)

In 1990 Herman sky developed the Perceptual Linear Prediction (PLP) model. The objective of the unique perceptual linear prediction model is to explain the psychophysics of individual hearing more precisely in the feature extraction process. Perceptual linear prediction rejects inappropriate information of the language and hence improves speech recognition rate.

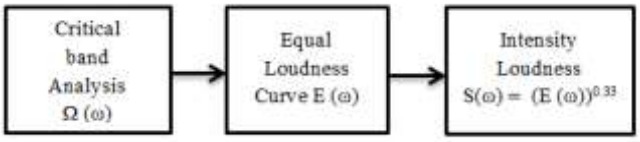

Fig.4. Block diagram of PLP Processing

Fig. 4 illustrates stages of perceptual linear prediction computation which estimates three major perceptual features as shown in above block diagram and the three features are identified as the cubic root.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

Figure 5 shows detailed steps of PLP computation. From signal that is windowed, power spectrum can be determined as:

$$P(\omega) = Re(S(\omega))^2 + Im(S(\omega))^2 \quad (1)$$

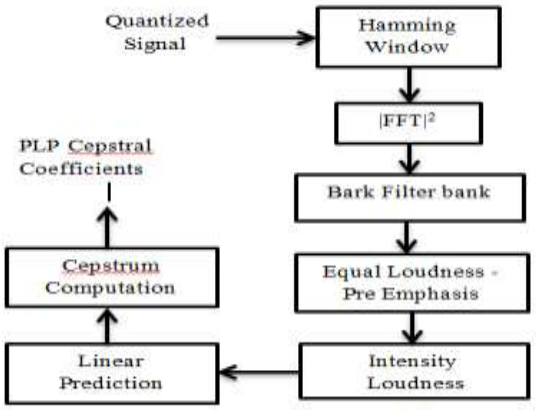

FIG. 5. PLP PARAMETER COMPUTION

The frequency warping into the Bark scale is applied. Conversion from frequency to bark is the first step, which is a better representation of the individual hearing resolution in frequency. The bark frequency that corresponds to an audio frequency is

$$\Omega(\omega) = 6ln\left(\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right) \quad \textbf{(2)}$$

III. PERFORMANCE EVALUATION

The Performance of speech recognition system can be measured. Accuracy and speed are two most frequently used performance measurements. Word Error Rate (WER) is used to measure accuracy, while real time factor is used to measure speed. We can calculate WER by using the equation (1)

$$WER = S + D + I/N \quad (1)$$

In which,
S implies number of substitutions,
D implies number of the deletions,
I implies number of the insertions and
N implies number of words in the reference.

Real Time Factor (RTF) is used to measure the speed of a speech recognition system. Let P be the time taken to process an input of duration I then RTF is given as,

$$RTF = P/I \quad (2)$$

IV. CONCLUSION

In this paper a brief description of Automatic Speech Recognition is provided and three different feature extraction methods for speech recognition systems are introduced.

Mel Frequency Cepstral Coefficient's (MFCC) is for the most part repeatedly applied feature extraction technique in speech recognition systems because it is most nearby to the actual individual hearing speech perception.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

V. REFERENCES

[1] Anjivani S. Bhabad, Gajanan K. Kharate, An Overview of Technical Progress in Speech Recognition, International Journal of Advanced Research in Computer Science and Software Engineering, March 2013,.Volume 3, Issue 3.
[2] M.A.Anusuya, S.K.Katti, "Speech Recognition by Machine: A Review", International Journal of Computer Science and Information Security, Vol. 6, No. 3, pp. 181-205, 2009
[3] Om Prakash Prabhakar, Navneet Kumar Sahu, "A Survey On: Voice Command Recognition Technique", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, pp. 576-585, May 2013
[4] Suman K. Saksamudre, P.P. Shrishrimal, R.R. Deshmukh,"A Review on Different Approaches for Speech Recognition System", International Journal of Computer Applications, Volume 115 – No. 22, pp. 23-28, April 2015
[5] Lei Xie, Zhi-Qiang Liu, "A Comparative Study of Audio Features For Audio to Visual Cobversion in MPEG-4 Compliant Facial Animation," Proc. of ICMLC, Dalian, 13-16 Aug-2006.
[6] Alfie Tan Kok Leong, "A Music Identification System Based on Audio Content Similarity," Thesis of Bachelor of Engineering, Division of Electrical Engineering, The School of Information Technology and Electrical Engineering, The University of Queensland, Queensland, Oct-2003.
[7] Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia.
[8] DOUGLAS O'SHAUGHNESSY, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", **Proceedings of the IEEE, VOL. 91, NO. 9, September 2003, 0018-9219/03 2003 IEEE**.
[9] N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, "A Hybrid model of Neural Network Approach for Speaker A.P.Henry Charles & G.Devaraj, "Alaigal-A Tamil Speech Recognition", Tamil Internet 2004, Singapore.
[10] independent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010 1793-8201.
[11] Xian Tang, "Hybrid Hidden Markov Model and Artificial Neural Network for Automatic Speech Recognition," Pacific-Asia Conference on Circuits, Communication and System, **IEEE Computer society, 2009.**
[12] Nidhi Desai, Prof. Kinnal Dhameliya, "Feature Extraction and Classification Techniques for Speech Recognition: A Review," International Journal of Emerging Technology and Advanced Engineering (IJETAE), Vol. 3, Issue 12, December 2013.
[13] Revathi, Y. Venkataramani, "Speaker Independent Continuous Speech and Isolated Digit Recognition using VQ and HMM," **IEEE conference, pp. 198-202, 2011.**
[14] P.G.N. Priyadarshani, N. G. J. Dias, AmalPunchihewa, "Dynamic Time Warping Based Speech Recognition for Isolated Sinhala Words," **IEEE Journal, pp. 892-895, 2012.**
[15] Ahmad A. M. Abushariah, Teddy S. Gunawan, Mohammad A. M. Abushariah, "English Digit Speech Recognition System Based on Hidden Markov Model," International Conference on Computer and Communication Engineering, **IEEE, May 2010.**
[16] C. Sunitha Ram, Dr. R. Ponnusamy, "An Effective Automatic Speech Emotion Recognition for Tamil Language using Support Vector Machine", International Conference on Issues and Challenges in Intelligent Computing Techniques, 2014.
[17] Huo Chun bao, Zhang Caijuan, "The Research of speaker recognition based on GMM and SVM", International Conference on System Science and Engineering, **IEEE, pp.373-375, July 2012.**
[18] Md. R. Hasan, M. Jamil, Md. G. Rabbani, Md. S. Rahman, "Speaker Identification using Mel Frequency Cepstral Coefficients", 3rd International Conference on Electrical & Computer Engineering, Dhaka, December 2004.
[19] Sanjivani S. Bhabad, Gajanan K. Kharate, "Overview of Technical Progress in Speech Recognition", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol.3, Issue 3, March 2013.
[20] Ranu Dixit, NavdeepKaur, "Speech Recognition Using Stochastic Approach: A Review", International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), Vol.2, Issue 2, February 2013.
[21] L. Muda, M. Begam, I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Vol.2, Issue 3, March 2010.
[22] Santosh K. Gaikwad, Bharti W. Gawali, PravinYannawar, "A Review on Speech Recognition Technique," International Journal of Computer Applications (IJCA), Vol. 10, Issue 3, Nov. 2010.
[23] Divyesh S. Mistry, Prof. A.V. Kulkarni, "Overview: Speech Recognition Technology, Mel-Frequency Cepstral Coefficients (MFCC), Artificial Neural Network (ANN)", International Journal of Engineering Research & Technology (IJERT), Vol.2, Issue 10, October 2013.[23]
[24] Goh Kia Eng, Abdul Manan Ahmad, "Malay Speech Recognition using Self-Organizing Map and Multilayer Perceptron", Proceeding of the Postgraduate Annual Research Seminar, 2005.
[25] Bishnu Prasad Das, Ranjan Parekh, "Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE with Neural Network Classifiers," International Journal of Modern Engineering Research,(IJMER) Vol. 2, Issue 3,pp.854-858, June 2012.

[26] A. Hmich, A. Badri, A. Sahel, "Automatic Speaker Identification by using Neural Network", **IEEE conference, 2010**.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

[27] S "DWT and MFCC Based Human Emotional Speech Classification Using LDA" International Conference on Biomedical Engineering (ICoBE), Penang, 27-28 February 2012, pp. 203-206.
[28] Michael Pitz, Ralf Schl¨uter, and Hermann Ney Sirko Molau, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum," in 2001 **IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01), USA, 2001, pp. 73-76.**
[29] Areg G. Baghdasaryan and A. A. (Louis) Beex "Automatic Phoneme Recognition with Segmental Hidden Markov Models" **IEEE 2011 Conference on Signals, Systems and Computers, ASILOMAR, 2011, pp. 569-574.**
[30] Han, W., Chan, C.F., Choy, C.S., et al.: 'An efficient MFCC extraction method in speech recognition'. **The 2006 IEEE Int. Symp. on Circuits and Systems, 2006, pp. 145–148**
[31] Lawrence Rabiner, and Biing Hwang Juang, Fundamentals of Speech Recognition. Prentice Hall, New Jersey, 1993.
[32] Ethnicity Group. "Cepstrum Method". 1998 http://www.owlnet.rice.edu/~elec532/PROJECTS98/speech/cepstrum/cepstrum.html
[33] Thiang and Suryo Wijoy "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot" **The 2011 International Conference on Information and Electronics Engineering IPCSIT vol.6 (2011) © (2011) IACSIT Press, Singapore**
[34] Lei Xie, Zhi-Qiang Liu, "A Comparative Study of Audio Features For Audio to Visual Cobversion in MPEG-4 Compliant Facial Animation," Proc. of ICMLC, Dalian, 13-16 Aug-2006.

CITE AN ARTICLE

Jolad, B., & Khanai, R., Dr. (n.d.). DIFFERENT FEATURE EXTRACTION TECHNIQUES FOR AUTOMATIC SPEECH RECOGNITION: A REVIEW. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *7*(2), 181-188.