Fourier Transform and Correlation-based Feature Selection for Fault Detection of Automobile Engines

Hamid Ghaderi
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
h_ghaderi@comp.iust.ac.ir

Peyman Kabiri
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
peyman.kabiri@iust.ac.ir

***Abstract*—Recently, research on effective Acoustic Emission (AE)-based methods for condition monitoring and fault detection has attracted many researchers. Due to the complex properties of acoustic signals, effective features for fault detection cannot be easily extracted from the raw acoustic signals. To extract representative features, signal processing techniques play an important role. One of the commonest techniques is Fast Fourier Transform (FFT). This method depends on the variations in frequency domain to distinguish different operating conditions of a machine. In this study, the intension is to categorize the acoustic signals into healthy and faulty classes. Acoustic emission signals are generated from four different automobile engines in both healthy and faulty conditions. The investigated fault is within the ignition system of the engines while they might suffer from other possible problems as well that may affect the generated acoustic signals. The energy of FFT coefficients of acoustic signals for different frequency bands are calculated as features. Correlation-based Feature Selection (CFS) algorithm is used to reduce the dimensionality of the dataset. The case study is carried-out on 4 different types of automobiles using 480 automobiles to prove the independency of the proposed approach on the type of the automobile. Classification results are reported to be around 88 percent accuracy.**

Keywords-Acoustic Emission (AE); Fast Fourier Transform (FFT); fault detection; condition monitoring; Correlation-based Feature Selection (CFS)

I. INTRODUCTION

Rapid automobile industry growth has made engine's maintenance to be of great importance. Therefore, it seems necessary to development accurate condition monitoring and fault detection systems for both reducing maintenance cost and alerting the operator about the engine's operating condition before severe damages occur. Stress wave travels through the materials and is caused by sudden release of strain energy. This stress wave is called an Acoustic Emission (AE) wave [1]. AE as a non-destructive testing method has been widely used by a lot of researchers in many industries. For instance, fault detection and condition monitoring of mechanical components such as gearboxes [2], engines [3] and bearings [4] have been the target of AE based methodologies. Fortunately, the operating condition of such components can be monitored by their dynamic information that is present in AE wave forms emitted from them. Internal Combustion (IC) engines are typical types of rotating machineries. Fault diagnosis and condition monitoring of such engines using acoustic signals have been the target of a lot of research projects. Wu and Chuang [5] have investigated cooling fan and drive axel shaft faults of vehicles with four cylinder IC engines. Using visual dot pattern technique along with acoustic and vibration signals, they have produced a snowflake-shaped pattern of six fold symmetry. Their proposed fault diagnosis procedure is completed by adopting an automatic image template matching. In another work, Kabiri and Makinejad [6] have investigated the combustion fault in Pride automobile (Kia motors). They have used Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT) for features extraction from acoustic signals. Jiang et al. [7] have focused on condition monitoring of four cylinder diesel engine with combustion faults using acoustic measurements. Using one-port acoustic source theory, they have concluded that a better representation of engine combustion condition is obtained by the strength of engine acoustic source. Wu and Chen [8] used Continuous Wavelet Transform (CWT) for both vibration and acoustic based fault diagnosis of two experimental works: IC engine and its cooling fan blade defects. Wu and Liu [9] have investigated the fault diagnosis process of IC engine with different faults using DWT and neural networks.

One of the most significant issues is how to extract relevant features from acoustic signals to help fault detection and condition monitoring of those engines to be carried out as accurately as possible. This issue is highly dependent on the appropriate signal processing technique used for feature extraction. Among many signal processing techniques used in the literature, FFT is one of the most popular ones and it is greatly utilized in condition monitoring and fault diagnosis [10]. FFT is a frequency domain analysis that is used to extract frequency domain features [11]. This method relies on the variations in frequency to isolate various faulty conditions. FFT transfers signals to the frequency domain, a process that results in using only frequency domain information regardless of time domain information.

In this paper, acoustic signals of four different engines in both healthy and faulty operating conditions are recorded and analyzed using FFT. Spectrum of the signals is divided into different frequency segments. The energy is calculated as a feature using FFT coefficients in each frequency segment. As

978-1-4673-1479-4/12/$31.00 ©2012 IEEE

the datasets may include irrelevant features that may affect the classification accuracy, feature reduction is needed. Feature selection algorithms not only select the more relevant features but also reduce volume of the dataset. For this purpose, the Correlation-based Feature Selection (CFS) algorithm is adopted. The dataset with reduced dimensionality is then classified using Support Vector Machine (SVM) with Radial Basis Function as the kernel function, K-Nearest Neighbor (KNN) with parameter K=5 and Multi-Layer Perceptron (MLP) with Back Propagation as the learning algorithms. It should be mentioned that the classification accuracy is validated and reported using stratified 10-fold cross validation in which 10 percent of data is randomly selected for training and 90 percent for testing. The classification results show efficiency of the FFT-based feature extraction for the reported case study in this paper.

II. FOURIER TRANSFORM

An energy-limited signal *f(t)* can be decomposed by its Fourier transform *F(ω)*, namely

$$f(t)=\int_{-\infty}^{+\infty}F(\omega)e^{iwt}d\omega \tag{1}$$

Where

$$F(\omega)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}f(t)e^{-iwt}dt \tag{2}$$

f(t) and *F(ω)* are a pair of Fourier transforms. Equation (1) implies that *f(t)* signal can be decomposed into a group with harmonics $e^{i\omega t}$. The weighting coefficients *F(ω)* represent the amplitudes of the harmonics in *f(t)*. *F(ω)* is time independent and it represents the frequency composition of a random process, which is assumed that its statistics do not change with time.

III. FEATURE EXTRACTION

In this paper, the frequency spectrums of signals are segmented into 9 different bands including 50 Hz, 100 Hz, 250 Hz, 500 Hz, 1000 Hz, 1500 Hz, 2000 Hz, 3000 Hz, and 5000 Hz. Fig. 1 shows the 2500 Hz segmentation of frequency spectrum of a signal. As the faults affect the signals of normal condition in the frequency domain, the aim is to find the best frequency segment where the fault has affected the signals significantly. On the other hand, the frequency segmentation resolution influences the number of features extracted from the spectrum of the signals. Frequency segmentation resolution represents the precision of segmentation. For example the 50Hz frequency segmentation has more segmentation resolution than the 1000Hz frequency segmentation i.e. the 50Hz segmentation has focused on the spectrum of the signal in more detail. There is a kind of trade-off between the frequency segmentation resolution and the number of features.

For each band, the energy of the absolute value of FFT coefficients is calculated as a feature i.e. x_i in the energy formulations that is shown in (3).

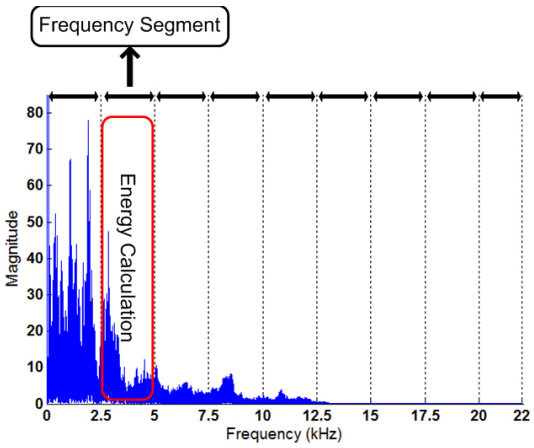

Figure 1. Frequency spectrum segmentation of a signal

$$Energy=\sum_{i=1}^{N}|x_i|^2 \tag{3}$$

IV. CORRELATION BASED FEATURE SELECTION (CFS)

CFS algorithm tries to find a subset of features not only to lower the dimensionality of the dataset but also to improve the classification accuracy. CFS defines a merit for each selected subset of features. The merit is based on this hypothesis that a promising subset of features involves those features that are uncorrelated or less correlated to each other while they are highly correlated to the class label. The merit is mathematically defined as (4) [12]:

$$Merit_S=\frac{k\overline{r_{c,f}}}{\sqrt{k+k(k-1)\overline{r_{f,f}}}} \tag{4}$$

Where $Merit_S$ indicates that how worthy the feature subset *S* with *k* features is. Parameters $\overline{r_{c,f}}$ and $\overline{r_{f,f}}$ are mean feature-class and mean feature-feature correlations respectively, where $f \in S$. The predictive ability of feature subset *S* and the amount of redundancy among its features are calculated by the nominator and denominator of (4). The more the features are correlated to each other, the more redundancy there is among them. Thus $Merit_S$ of a feature subset has a smaller value than the time when the features in the subset are uncorrelated with each other. This should be mentioned that the numerical features are to be discritized before CFS is applied. Discretization is the process of transforming continuous valued attributes to nominal ones. Here the supervised discretization technique proposed by Fayyad and Irani [13] is utilized which is described in section V.

Symmetric uncertainty shows the correlation between two features *X* and *Y* that is presented as (10) [12]:

$$\text{Symmetric uncertainty}=\frac{2\times gain}{H(X)+H(Y)} \tag{10}$$

Where *H(.)* is the entropy of the feature and *gain* indicates the information gain. Entropy is considered to be a measure of

uncertainty or unpredictability in a system. The entropy of feature X is calculated by (11).

$$H(X) = -\sum_{x \in X} p(x) \log_2 \left(p(x) \right) \tag{11}$$

Where, $p(x)$ is the probability of nominal values for feature X ($x \square X$).

Information gain is the amount of information gained about Y after observing X. Information gain is a symmetric measure. Equation (12) [14] represents the information gain.

$$gain = H(X) + H(Y) - H(X,Y) \tag{12}$$

Before applying CFS to reduce the dataset dimensionality, features are normalized. If the feature space has n dimensions, the number of the possible feature subsets will be 2^n. Therefore, it seems necessary to use a certain search strategy to explore the feature space. Best First Search (BFS) strategy has been used in this paper. Starting from an empty subset of features, BFS tries to explore the feature space by making the local changes to the current feature subset. In BFS, unlike the greedy hill climbing, back tracking is allowed. That is, while exploring the feature space, if local changes to the current feature subset begin to look less promising, BFS back tracks to a more promising previous subset of features and continues the search from there. Technically, without any stopping criterion, BFS explores the entire feature space. To avoid it in this paper, the number of fully expanded subsets that lead to no improvement is limited to 5. The BFS algorithm is shown in the Fig. 2.

V. FEATURE DISCRETIZATION

Based on minimum entropy heuristic, the utilized discritization technique uses the class information entropy of candidate partitions to select the most promising partition boundary for discritization. The partition boundary is called cut point and is referred to as T. For a given subset of instances (S) and a feature (A), T partitions S into 2 subsets (S_1 and S_2). The class information entropy of S_1 and S_2 is calculated as (5):

$$E(A,T;S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \tag{5}$$

Where $Ent(S)$ is the class entropy of S. $Ent(S)$ is calculated as (6):

$$Ent(S) = -\sum_{i=1}^{C} p(C_i, S) \log_2 (p(C_i, S)) \tag{6}$$

Where the probability of the class C_i represented in S is shown as $p(C_i, S)$.

A cut point T is selected if and only if the following condition is met.

$$Gain(A,T;S) > \frac{\log_2 (N-1) +}{N} + \frac{\Delta(A,T;S)}{N} \tag{7}$$

$Gain(A,T; S)$ and $\Delta(A,T; S)$ are calculated as (8) and (9), respectively. N is the number of instances in S.

$$Gain(A,T;S) = Ent(S) - E(A,T;S) \tag{8}$$

$$\Delta(A,T;S) = \log_2 (3^k - 2) - \begin{bmatrix} kEnt(S) - k_1 Ent(S_1) \\ -k_2 Ent(S_2) \end{bmatrix} \tag{9}$$

In (9), k, k_1, and k_2 are the number of different classes in S, S_1, and S_2, respectively.

The above procedure is continued considering each S_1 and S_2 as S till no more cut points are found.

1. Begin with the OPEN list containing the start state, the CLOSED list empty, BEST← start state.
2. Let s = arg max e(x) (get the state from OPEN with the highest evaluation).
3. Remove s from OPEN and add it to CLOSED.
4. If e(s) ≥ e(BEST), then BEST← s.
5. For each child t of s that is not in the OPEN or CLOSED list, evaluate and add to OPEN.
6. If BEST changed in the last set of expansions, go to 2.
7. Return BEST.

Figure 2. Best First Search (BFS) Algorithm

VI. EXPERIMENTS

The investigated engine defect is in the ignition system i.e. engines operate with one cylinder missing fire. In the reported work, the spark in the first cylinder is not happening. The engine acoustic signals are recorded in the workshop using a microphone 20 cm above the engine. The investigated engines are from 4 different automobiles including Pride (Kia motors), Peugeot 405, Peugeot Pars, and Iranian national automobile Samand. For each automobile, the acoustic signals of 60 different automobiles in both healthy and faulty conditions with engines operating with 1000 rpm and 44100 sampling frequency are recorded in WAV format. The recording process is carried-out in the central repair shop of the Iran Khodro car manufacturing company.

A. Pre-processing

The recorded signals have been manually de-noised. Due to recording the signals in workshop, the sound of other working objects near the test subject and human voice is considered noise. As in the workshop the automobiles were to be checked by the repairman, the automobiles may or may not suffer from other possible faults. For example, the combustion timing defect causes the engine not to operate properly and this per se results in considerable acoustic abnormality. The analyzed signals include both healthy and faulty operating conditions with the recording time of 5 seconds.

B. Classification train and test datasets

The total number of recorded signals is a dataset with 480 samples including 60 samples for each type of automobile with 2 classes: healthy and faulty. The dataset is divided into train and test datasets. For the training dataset, 10 percent of data is randomly selected by stratified 10-fold cross validation strategy

and 90 percent remained as the testing dataset. The aim of selecting only 10 percent of data for training is to prove the generalization capability of the proposed method.

VII. RESULTS

As the recording sampling rate is 44100 samples per second, the covered frequency range is [0-22050] Hz. The frequency spectrum of each signal is segmented into several frequency bands. This frequency bands can be referred to as frequency segmentation resolutions. The segmentation resolutions are: 50 Hz, 100 Hz, 250 Hz, 500 Hz, 1000 Hz, 1500 Hz, 2000 Hz, 3000 Hz, and 5000 Hz. For each segment, the aforementioned feature is calculated using the absolute value of FFT coefficients.

The number of features extracted from each signal using a frequency segmentation resolution of N Hz is calculated by (13).

$$No.\ of\ \ Features = \left\lceil \frac{22050}{N} \right\rceil \tag{13}$$

The constructed datasets for each frequency segmentation resolution are normalized first and then without any dimensionality reduction they were used for classification. TABLE I shows number of the extracted features for each dataset and the classification results using those datasets. Classification results are presented in terms of the Accuracy (ACC), True Positive Rate (TPR) and False Positive Rate (FPR).

To look closer at TABLE I it is obvious that the number of extracted features especially for datasets with low segmentation resolution is unnecessarily high. After applying CFS on the primary datasets and then normalizing their features for classification TABLE II is obtained. TABLE III shows the selected features from each constructed dataset after applying CFS and the approximated covered frequency.

To use FFT, it is considered that the original signals are stationary i.e. the statistical characteristics of the signals do not change over time. However, the acoustic signals emitted from the automobile engines are non-stationary [15]. Furthermore, applying Fourier transform over a non-stationary signal generally equates for a frequency composition averaged over the duration of the signal. This results for that the Fourier transform would not be able to adequately describe the characteristics of the signal in low frequencies. In other words, the proposed methodology concentrates on the high frequency components of the signal and tries to find an accurate signature for the investigated fault. This is proved by TABLE III where the selected features are mainly from high frequencies.

Comparing TABLE I with TABLE II reveals that the CFS considerably improves the classification results. CFS also reduces the required number of features for dataset classification remarkably. Furthermore, the best results are obtained by 1500 Hz frequency spectrum segmentation resolution with the accuracy of 87.19 percent using SVM for classification.

VIII. CONCLUSION

The reported work uses AE signal analysis to identify faulty combustion of an automobile engine regardless of the type of automobile. The analyzed AE signals are recorded in the workshop with the presence of environmental noise, yet the proposed methodology can still operate. As the signals are recorded from four different types of automobile engines, one can claim that the methodology used in this paper has this potential to be used for different types of automobile engines. Therefore, it is possible to say that the proposed method is automobile independent. Considering the reported result, suitability of FFT based features as well as CFS algorithm for feature selection is proved. The generalization capability of the proposed methodology is proved using only 10 percent of data for training and 90 percent for testing.

IX. FUTURE WORKS

Intention is to improve the classification results using more appropriate feature extraction and signal processing techniques. For example, using the time-frequency transforms to use both time and frequency characteristics of signals are in mind. At the same time, detection of the automobile in the form of specific automobile detection or categorized detection of similar automobiles are considered. Adding more faults to the list of the faults and successful classification of them is also included in the future plan for the reported work. Publicizing the signals used in this paper is expected via our laboratory website: http://ial.iust.ac.ir/.

ACKNOLEDGMENTS

Authors' thank are to Irankhodro Powertrain COmpany (IPCO) a subsidiary of Iran Khodro Company a leading Iranian automaker (Mr. Izanloo) and Iran Khodro central repair shop number 5 (Mr. Saghi) that supported this work by giving us access to their facilities to collect samples.

REFERENCES

[1] X. Li, "A brief review: acoustic emission method for tool wear monitoring during turning", International Journal of Machine Tools & Manufacture, vol. 42, pp. 157–165, January 2002.

[2] B B. Eftekharnejad, and D. Mba, "Seeded fault detection on helical gears with acoustic emission", Applied Acoustics, vol. 70, pp. 547-555, April 2009.

[3] A. Albarbar, F. Gu, and A. D. Ball, "Diesel engine fuel injection monitoring using acoustic measurements and independent component analysis", Measurement, vol. 43, pp. 1376-1386, 2010.

[4] A. M. Al-Ghamd, and D. Mba, "A comparative experimental study on the use of acoustic emission and vibration analysis for bearing defect identification and estimation of defect size", Mechanical Systems and Signal Processing, vol. 20, pp. 1537-1571, October 2006.

[5] J. D. Wu, and C. Q. Chuang, "Fault diagnosis of internal combustion engines using visual dot patterns of acoustic and vibration signals", NDT & E International, vol. 38, pp. 605-614, December 2005.

[6] P. Kabiri, and A. Makinejad, "Using PCA in Acoustic Emission Condition Monitoring to Detect Faults in an Automobile Engine." presented in 29th European Conference on Acoustic Emission Testing (EWGAE2010), September 2011.

[7] J. Jiang, F. Gu, R. Gennish, D. J. Moore, G. Harris, and A. D. Ball, "Monitoring of diesel engine combustions based on the acoustic source characterisation of the exhaust system", Mechanical Systems and Signal Processing, vol. 22, pp. 1465-1480, August 2008.

TABLE I. THE CLASSIFICATION RESULTS USING THE CONSTRUCTED DATASET FOR EACH SEGMENTATION RESOLUTION BEFORE FEATURE SELECTION

Segmentation Resolution (Hz)	No. of Features	SVM			KNN			MLP		
		ACC (%)	*TPR (%)*	*FPR (%)*	*ACC (%)*	*TPR (%)*	*FPR (%)*	*ACC (%)*	*TPR (%)*	*FPR (%)*
50	441	57.94	57.51	40.63	52.73	54.53	48.6	-----	-----	-----
100	222	61.37	64.17	40.52	55.9	52.99	41.14	-----	-----	-----
250	89	67.45	76.28	40.97	60.18	83.3	37.3	64.5	63.3	34.1
500	45	73.45	80.11	33.17	77.75	69.94	14.29	68.89	84.25	46.48
1000	23	70.09	68.81	28.29	77.15	72.51	18.04	69.37	75.24	35.89
1500	16	71.06	63.21	20.8	78.52	73.43	16.37	72.89	76.3	30.11
2000	12	77.64	76	21.31	78.17	75.2	18.65	73.54	67.5	20.52
3000	8	77.38	74.45	19.65	78.47	74.73	17.56	76.97	76.49	22.16
5000	5	81.85	84.05	20.18	79.26	79.43	20.89	72.82	76.45	30.59

TABLE II. THE CLASSIFICATION RESULTS USING THE CONSTRUCTED DATASET FOR EACH SEGMENTATION RESOLUTION AFTER FEATURE SELECTION

Segmentation Resolution (Hz)	No. of Features	SVM			KNN			MLP		
		ACC (%)	*TPR (%)*	*FPR (%)*	*ACC (%)*	*TPR (%)*	*FPR (%)*	*ACC (%)*	*TPR (%)*	*FPR (%)*
50	24	74.82	78.89	28.97	75.54	63.77	12.5	68.12	62.67	26.23
100	19	80.56	79	17.41	77.07	64.44	10.05	69.44	61.81	22.81
250	9	83.45	83.1	16.02	81.39	74.26	11.29	74.22	71.54	23.07
500	7	84.45	81.39	12.32	82.75	75.15	9.48	75.41	66.94	15.94
1000	5	85.18	83.04	12.52	84.43	77.64	8.71	76.53	67.13	14.11
1500	4	87.19	84.77	10.14	87.83	85.34	9.6	78.72	78.92	21.26
2000	3	84.07	80.17	11.9	84.46	79.76	10.82	74.51	63.25	14.24
3000	3	82.44	79.08	14.02	80.86	72.56	10.75	74.12	62.16	13.68
5000	3	82.49	80.13	14.89	80.44	71.33	10.21	75.07	67.64	17.33

TABLE III. SELECTED FETURES FOR EACH SEGMENTATION RESOLUTION AND THE APPROXIMATED COVERED FREQUENCY BAND

Segmentation Resolution (Hz)	Selected Features (index of features)	Covered Frequency Range (Approximated)
50	13 56 166 169 196 202 216 221 234 236 251 254 261 350 377 378 383 389 392 400 404 406 407 408	(600~650) Hz (2750~2800) Hz (8250~13050) Hz (17450~20400) Hz
100	83 86 98 108 111 114 121 126 128 131 188 190 192 196 197 200 202 203 204	(8200~13100) Hz (18700~20400) Hz
250	34 43 48 51 76 77 79 81 82	(8250~12750) Hz (18750~20500) Hz
500	22 25 26 36 38 40 41	(10500~13000) Hz (18000~22050) Hz
1000	12 13 19 20 21	(11000~13000) Hz (18000~22050) Hz
1500	8 9 13 14	(10500~13500) Hz (18000~22050) Hz
2000	7 10 11	(12000~14000) Hz (18000~22000) Hz
3000	4 7 8	(9000~12000) Hz (18000~22050) Hz
5000	3 4 5	(10000~22050) Hz

[8] J. D. Wu, and J. C. Chen, "Continuous wavelet transform technique for fault signal diagnosis of internal combustion engines", NDT & E International, vol. 39, pp. 304-311, June 2006.

[9] J. D. Wu, and C. H. Liu, "Investigation of engine fault diagnosis using discrete wavelet transform and neural network", Expert Systems with Applications, vol. 35, pp. 1200-1213, October 2008.

[10] G. Betta, C. Liguori, A. Paolillo, and A. Pietrosanto, "A DSP-based FFT-analyzer for the fault diagnosis of rotating machine based on vibration analysis", IEEE Transactions on Instrumentation and Measurement, vol. 51, pp. 1316-1322, December 2002.

[11] M. E. H. Benbouzid, "A review of induction motors signature analysis as a medium for faults detection", IEEE Transactions on Industrial Electronics, vol. 47, pp. 984 - 993, Octobr 2000.

[12] M. A. Hall, and L. A. Smith, "Feature Selection for Machine Learning: Comparing Correlation-based Filter Approach to the Wrapper." in Proc. 12th International Florida Artificial Intelligence Research Society Conference (FLAIRS-99) 1999, pp. 235-239.

[13] U. M. Fayyad, and K. B. Irani, "Multi-interval discretization of continuous valued attributes for classification learning." in Proc. 13th International Joint Conference on Artificial Intelligence (IJCAI) 1993, pp. 1022-1027.

[14] J. R. Quinlan, “Induction of decision trees”, Machine Learning, vol. 1, pp. 81-106, 1986.

[15] F. Elamin, F. Gu, and A. Ball, “Diesel Engine Injector Faults Detection Using Acoustic Emissions Technique”, Modern Applied Science, vol. 4, pp. 3-13, September 2010.