Journal of Southeast University (English Edition) Vol. 25, No. 3, pp. 346 – 350 Sept. 2009 ISSN 1003—7985

Sparseness-controlled non-negative tensor factorization and its application in machinery fault diagnosis

Peng Sen Xu Feiyun Jia Minping Hu Jianzhong

(School of Mechanical Engineering, Southeast University, Nanjing 211189, China)

Abstract: Aiming at the problems of bispectral analysis when applied to machinery fault diagnosis, a machinery fault feature extraction method based on sparseness-controlled non-negative tensor factorization (SNTF) is proposed. First, a non-negative tensor factorization (NTF) algorithm is improved by imposing sparseness constraints on it. Secondly, the bispectral images of mechanical signals are obtained and stacked to form a third-order tensor. Thirdly, the improved algorithm is used to extract features, which are represented by a series of basis images from this tensor. Finally, coefficients indicating these basis images' weights in constituting original bispectral images are calculated for fault classification. Experiments on fault diagnosis of gearboxes show that the extracted features can not only reveal some nonlinear characteristics of the system, but also have intuitive meanings with regard to fault characteristic frequencies. These features provide great convenience for the interpretation of the relationships between machinery faults and corresponding bispectra.

Key words: non-negative tensor factorization; sparseness; feature extraction; bispectrum; gearbox

In machinery fault diagnosis, conventional analytical methods, such as the power spectrum, cannot effectively extract nonlinear fault features[1]. Besides, these methods lose phase information among frequencies and are not suitable for handling non-minimum phase systems or non-Gaussian signals[2]. Therefore, higher-order spectral analysis (HOSA) is usually used to extract and analyze the fault features of mechanical systems because a higher-order cumulant is not sensitive to additive Gaussian noise and symmetric non-Gaussian noise, and a higher-order spectrum can detect the phase coupling phenomenon[3] in signals. However, since the relationship between fault characteristic frequencies and higher-order spectra is less interpretable than that between fault characteristic frequencies and the second-order statistical quantities such as power spectra, traditional feature extraction methods still predominate over other methods in the practice of machinery fault diagnosis. In this paper we are concerned with bispectra, which is a simple case of higher-order spectra.

Non-negative tensor factorization (NTF) can compress and generalize large-scale high-dimensional data by factorizing them into a low-dimensional space and then extracting features from them. In 2001, Welling et al.[4] first proposed an algorithm for positive tensor factorization (PTF) and concluded that factors are easier to interpret than those produced by methods based on the singular value decomposition (SVD). In the following years, many other researchers also developed different NTF algorithms and applied them to various fields[5-8]. In NTF, high-dimensional data, such as an image cube, are factorized directly and approximated by the sum of rank-1 non-negative tensors[9]. Besides, NTF provides a unique factorization and its resulting "factors" are both sparse and separable[10]. This paper proposes a novel method using sparseness-controlled NTF (SNTF) to extract features from the bispectra of machinery fault signals.

1 Sparseness-Controlled NTF (SNTF)

The tensor factorization model for a third-order tensor $\boldsymbol{G}$ of dimensions $d_1 \times d_2 \times d_3$ can be expressed as

$$\boldsymbol{G} = \sum_{j=1}^{k} \boldsymbol{u}^j \circ \boldsymbol{v}^j \circ \boldsymbol{w}^j + \boldsymbol{E} \tag{1}$$

where $\boldsymbol{u}^j$, $\boldsymbol{v}^j$, $\boldsymbol{w}^j$ are the j-th column of matrices $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ ($\boldsymbol{u} \in \mathbf{R}^{d_1 \times k}$, $\boldsymbol{v} \in \mathbf{R}^{d_2 \times k}$, $\boldsymbol{w} \in \mathbf{R}^{d_3 \times k}$), respectively; "$\circ$" is a notation for the outer product of vectors, and $\boldsymbol{E}$ represents reconstruction or approximation error.

A typical NTF problem can be described as follows: Given a non-negative third-order tensor $\boldsymbol{G}$, find proper non-negative matrices $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ ($\boldsymbol{u} \in \mathbf{R}_+^{d_1 \times k}$, $\boldsymbol{v} \in \mathbf{R}_+^{d_2 \times k}$, $\boldsymbol{w} \in \mathbf{R}_+^{d_3 \times k}$) to minimize reconstruction error $\boldsymbol{E}$ in Eq. (1). To achieve this goal, consider solving the following least-squares problem:

$$\min_{\boldsymbol{u}^j, \boldsymbol{v}^j, \boldsymbol{w}^j} \left\| \boldsymbol{G} - \sum_{j=1}^{k} \boldsymbol{u}^j \circ \boldsymbol{v}^j \circ \boldsymbol{w}^j \right\|_{\mathrm{F}}^2 \tag{2}$$

$$\text{s. t.} \quad \boldsymbol{u}^j, \boldsymbol{v}^j, \boldsymbol{w}^j \geqslant \boldsymbol{0} \tag{3}$$

where $\| \boldsymbol{A} \|_{\mathrm{F}}^2$ stands for the Frobenius norm, i. e., the sum of squares of all the tensor elements $A_{r,s,t}$.

Based on the gradient descent scheme, Hazan et al.[10] proposed an update rule for $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$:

$$\left.\begin{aligned} u_i^m &\leftarrow \frac{u_i^m \sum_{s,t} G_{i,s,t} v_s^m w_t^m}{\sum_{j=1}^{k} u_i^j \langle \boldsymbol{v}^j, \boldsymbol{v}^m \rangle \langle \boldsymbol{w}^j, \boldsymbol{w}^m \rangle} \\ v_i^m &\leftarrow \frac{v_i^m \sum_{r,t} G_{r,i,t} u_r^m w_t^m}{\sum_{j=1}^{k} v_i^j \langle \boldsymbol{u}^j, \boldsymbol{u}^m \rangle \langle \boldsymbol{w}^j, \boldsymbol{w}^m \rangle} \\ w_i^m &\leftarrow \frac{w_i^m \sum_{r,s} G_{r,s,i} u_r^m v_s^m}{\sum_{j=1}^{k} w_i^j \langle \boldsymbol{u}^j, \boldsymbol{u}^m \rangle \langle \boldsymbol{v}^j, \boldsymbol{v}^m \rangle} \end{aligned}\right\} \tag{4}$$

Received 2009-04-21.

Biographies: Peng Sen (1986—), male, graduate; Xu Feiyun (corresponding author), male, doctor, professor, fyxu@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 50875048), the Natural Science Foundation of Jiangsu Province (No. BK2007115), the National High Technology Research and Development Program of China (863 Program) (No. 2007AA04Z421).

Citation: Peng Sen, Xu Feiyun, Jia Minping, et al. Sparseness-controlled non-negative tensor factorization and its application in machinery fault diagnosis[J]. Journal of Southeast University (English Edition), 2009, 25(3): 346 – 350.

where $\boldsymbol{u}^m$, $\boldsymbol{v}^m$, $\boldsymbol{w}^m$ are the m-th column of matrices $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$, respectively. Although this rule provides a unique factorization under weak constraints (There are only non-negativity constraints; see Eq. (3)), the sparseness of the factorization results cannot be controlled. So it is possible that the extracted features are not so sparse that they cannot reflect the local characteristics of original data(such as the data of an image cube). This drawback brings difficulty to the explanation of those extracted results.

Heiler et al. [11] proposed an approach based on the second-order cone programming (SOCP) to achieve precise sparseness control of NTF results. In this approach, sparseness control is converted into a conic optimization problem with the second-order conic constraints. Consider the sparseness control of $\boldsymbol{u}$ in Eq. (2):

$$\max_{\boldsymbol{u}, z} z \tag{5}$$

$$\text{s. t.} \quad f(\boldsymbol{G}, \boldsymbol{u}) \leqslant f(\boldsymbol{G}, \boldsymbol{u}^*) \tag{6}$$

$$z \leqslant \text{sp}(\boldsymbol{u}^{*j}) + \langle \nabla \text{sp}(\boldsymbol{u}^{*j}), \boldsymbol{u}^j - \boldsymbol{u}^{*j} \rangle \qquad j = 1, 2, \ldots, k \tag{7}$$

where $\boldsymbol{u}^j$ is the j-th column of $\boldsymbol{u}$, and $\boldsymbol{u}^{*j}$ is the estimate of $\boldsymbol{u}^j$ before sparseness optimization; $f(\boldsymbol{G}, \boldsymbol{u}) = \| \text{vec}(\boldsymbol{G}) - \boldsymbol{U}\text{vec}(\boldsymbol{u}) \|_2^2$ is the cost function or the reconstruction error function; $\text{vec}(\boldsymbol{G})$ is the vector form of tensor $\boldsymbol{G}$, and $\boldsymbol{U}$ is a sparse matrix containing two matrices except $\boldsymbol{u}$; and $\text{sp}(\boldsymbol{x})$ is vector $\boldsymbol{x}$'s sparseness measure, which was indicated in Ref. [11]. Nonetheless, this algorithm can optimize only one matrix in one iteration step, so the algorithm efficiency is not high.

In this paper, a sparseness-controlled algorithm for NTF is proposed by combining the advantages of the aforementioned algorithms. The algorithm is shown in algorithm 1($\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ are denoted as t_1, t_2, t_3 for the sake of convenience), where S_i is the smallest sparseness of the columns in t_i; $S_{i\min}$ is the lower bound of a desired sparseness; C_{old} and C_{new} are the reconstruction errors of two consecutive iterations and $C = \left\| \boldsymbol{G} - \sum_{j=1}^{k} \overset{3}{\underset{i=1}{\circ}} \boldsymbol{t}_i^j \right\|_{\text{F}}^2$. The upper bound of i in the algorithm is set to be 2.

Algorithm 1 SNTF

```
Initialize all t_i (i = 1, 2, 3) with arbitrary non-negative values
Iterate
  Calculate Eq. (4)
  for i = 1 to 2 do
    if S_i < S_imin
      Iterate
        Calculate Eqs. (5) to (7)
      Until the sparseness of t_i cannot be further improved
    else i = i + 1
  end for
Until | C_old - C_new | ≤ e
```

From the viewpoint of image feature extraction, the results of SNTF consist of two parts: k rank-1 matrices(or factors) $\boldsymbol{F}_j = \boldsymbol{u}^j \circ \boldsymbol{v}^j (1 \leqslant j \leqslant k)$, which form k basis images, each of which reflects a local characteristic extracted from a given image cube; and a $d_3 \times k$ matrix $\boldsymbol{w}$, which contains d_3 weight vectors with each vector involving k elements, each of which indicates a basis image's weight in constituting one of the d_3 images in the image cube.

2 Numerical Simulations

We first investigate the SNTF approach's capability in extracting features from a group of given bispectra, which are derived from a set of simulated amplitude modulation signals represented by

$$x(t) = A(1 + B\cos 2\pi f_{\text{n}} t)\sin(2\pi f_z t + \varphi) \tag{8}$$

where $A = 1$, $B = 1.5$, $f_z = 750$ Hz, $f_{\text{n}} = 25$ Hz and $\varphi = 0$. The sampling frequency f_{s} and the number of sampling points are set to be 4 000 Hz and 1 024, respectively.

By adding white noise to the existing sample in Eq. (8), 11 new samples are obtained and thus a sample set of 12 modulation signal samples is derived. Fig. 1 shows the bispectrum of the first sample of the simulated signal.

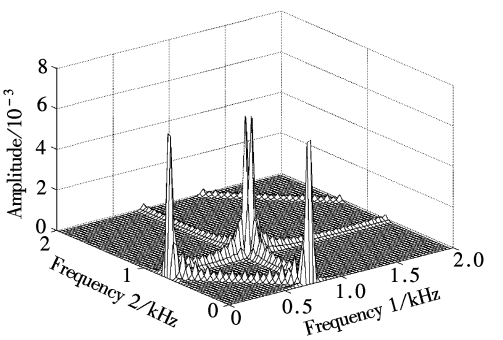

Fig. 1 Bispectrum of simulated amplitude modulation signal

After the bispectrum of each sample is obtained, an image cube consisting of 12 bispectral images is formed, the data form of which is a $64 \times 64 \times 12$ tensor, since each bispectral image(only the positive part is involved) is represented by a 64×64 matrix. Then different NTF methods are applied to the tensor. Figs. 2(a) to (c) show some basis images calculated from the three aforementioned NTF algorithms, Hazan's algorithm, Heiler's algorithm and the SNTF, respectively. Tab. 1 shows the performance comparison among the three algorithms.

Tab. 1 Performance comparison among the three algorithms

Algorithm	Reconstruction error/10^{-3}	Iteration steps
Hazan's algorithm	3.9	15
Heiler's algorithm	6.2	20
SNTF	3.4	10

From Fig. 2 and Tab. 1, it can be found that the basis images derived from both Heiler's algorithm and the SNTF are sparser than those derived from Hazan's algorithm, which has no sparseness constraints. However, Heiler's algorithm results in greater reconstruction errors and more iteration steps than the SNTF. So the SNTF is more suitable for extracting features from an image cube.

3 Experiments and Analysis

The test rigs are shown in Fig. 3, where three single-stage gearboxes(in the middle) are in three conditions—normal condition, a driving gear with pitting faults, and a driving gear with uniform wear. The number of teeth of the driving gears and driven gears are 31 and 46, respectively. Vibration

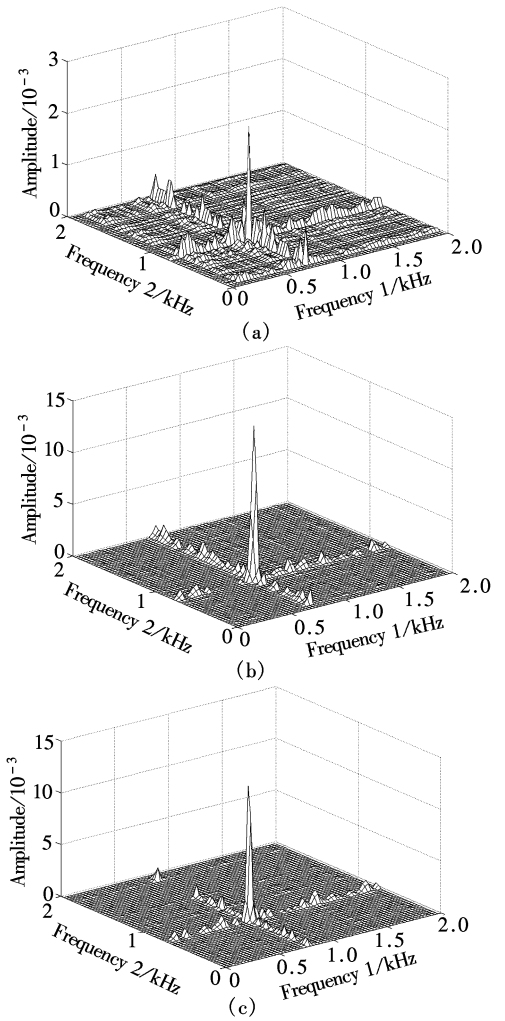

Fig. 2 Some basis images resulting from the three NTF algorithms. (a) Hazan's algorithm; (b) Heiler's algorithm; (c) SNTF

Fig. 3 The test rigs

signals are collected when the rotating speed of the driving gears is about 600 r/min. A total of 60 sample sets are collected from the three gearboxes (20 each), and each set is made up of 4 096 points. The meshing frequency of the gearboxes and the ball pass frequency of the outer races (BPFO) of bearings are 310 and 99.7 Hz, respectively. The sampling frequency is about 3 838 Hz. For the sample sets, 3/4 of them (15 sets for each state) are used for training and the remaining 1/4 (5 sets for each state) are used for testing. Fig. 4 shows the bispectra for the three states.

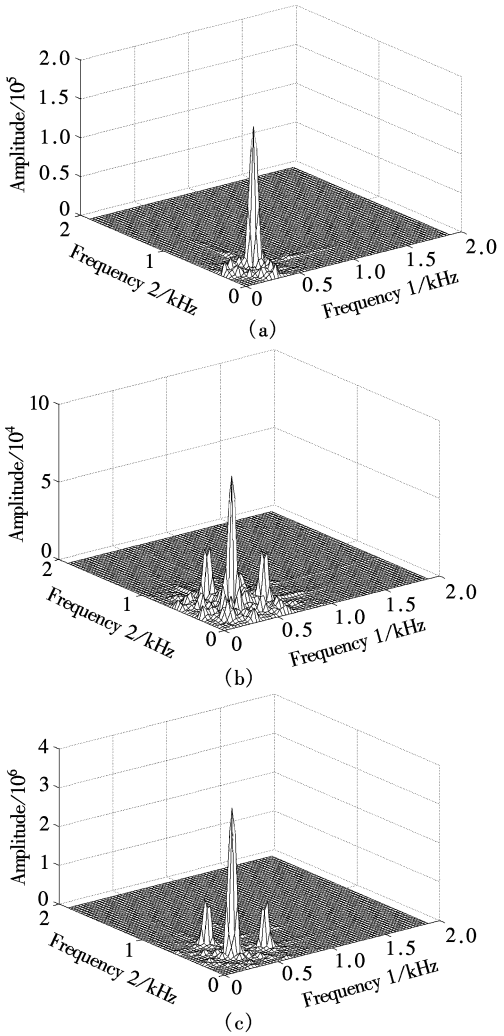

Fig. 4 Bispectrum of vibration signals in three states. (a) Normal; (b) Pitting; (c) Uniform wear

After the bispectrum of each sample is obtained, a training image cube that consists of 45 bispectrum images and a testing image cube that consists of 15 images are derived. Then the SNTF is applied to the training image cube. In the algorithm, we set $k = 20$. For basis images whose energies are localized in the same region, they are grouped and their sum is taken. Fig. 5 shows the resulting five superimposed new basis images.

From Figs. 5(a) to (c) it can easily be seen that the energy of each basis image is uniquely located around characteristic frequency pairs (310, 310 Hz), (620, 310 Hz), and(310, 620 Hz). So, compared with the bispectra in Fig. 4, the extracted features here carry more distinct meanings. Moreover, the experimental results reveal some features of quadratic nonlinearity. For example, the phenomenon of quadratic phase coupling is evident in Figs. 5(d) and (e), where energies are concentrated around frequency pairs (410, 210 Hz) and (210, 410 Hz), in which 210 and 410 Hz are approximate to the difference and the sum of the meshing frequency and BPFO, respectively. However, such a phenomenon cannot be seen in a bispectrum(see Fig. 4).

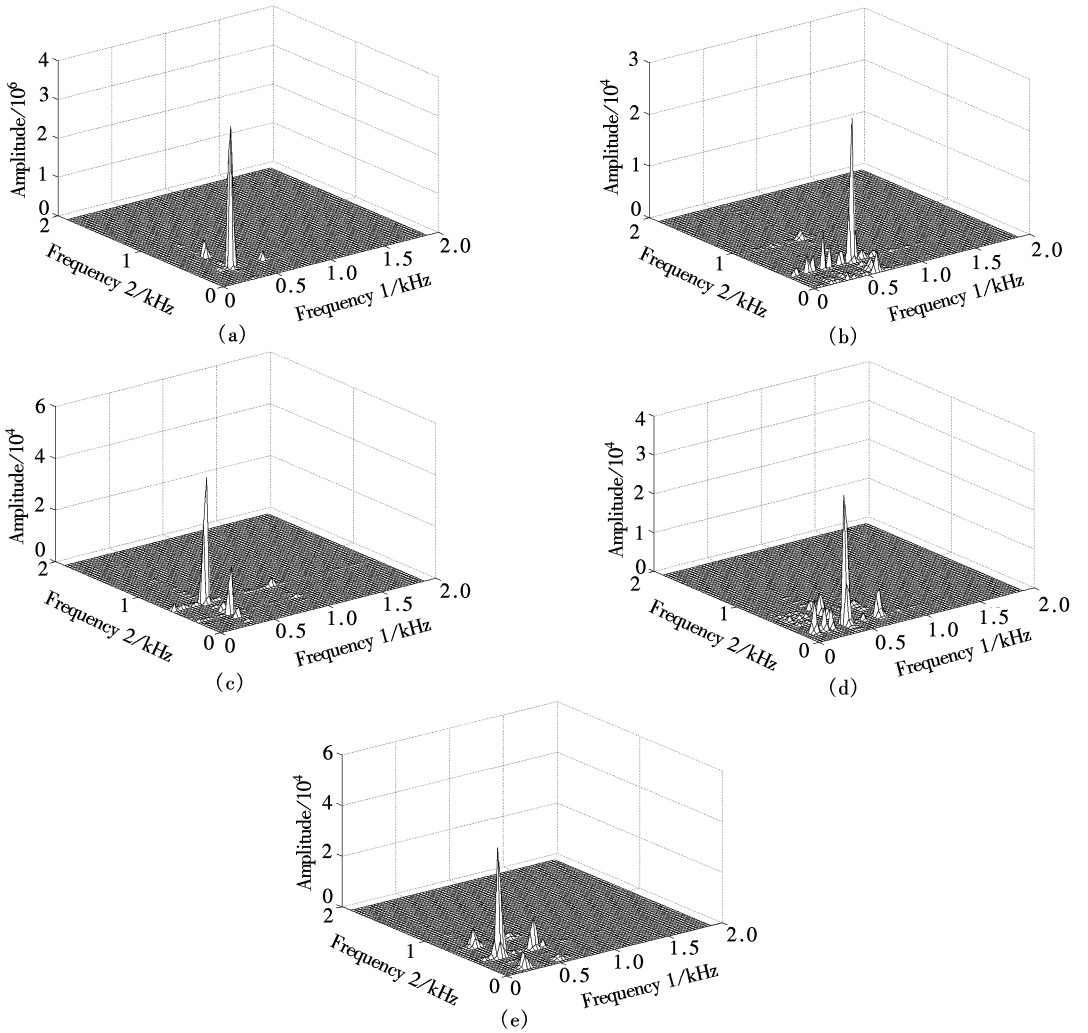

Fig. 5 Superimposed SNTF basis images

To find weight vectors denoting five basis images' weights in constructing a bispectral image in the given image cube, first all the basis images are normalized, and then a set of overdetermined equations is solved as follows:

$$\boldsymbol{b} = \boldsymbol{A}_1 x_1 + \boldsymbol{A}_2 x_2 + \ldots \boldsymbol{A}_j x_j + \ldots + \boldsymbol{A}_m x_m \qquad (9)$$

where $\boldsymbol{b}$ and $\boldsymbol{A}_j$ ($1 \leqslant j \leqslant m$, here $m = 5$) are matrices representing a certain original bispectrum image and a basis image, respectively; and x_j, an element of the weight vector $\boldsymbol{x}$, indicates basis image $\boldsymbol{A}_j$'s weight in constituting the image $\boldsymbol{b}$. Therefore, 45 training weight vectors and 15 testing weight vectors are obtained from the training and testing image cubes, respectively. Tab. 2 shows some of the training weight vectors out of the total of 45 (x_1 to x_5 correspond to basis images in Figs. 5(a) to (e), respectively).

Tab. 2 Some training weight vectors for the three states

Gearbox state	Weight of basis images				
	x_1	x_2	x_3	x_4	x_5
Normal	0.051 3	0.002 5	0.002 5	0.002 6	0.002 8
	0.053 8	0.001 1	0.002 7	0.002 3	0.002 6
	0.058 7	0.001 9	0.002 5	0.002 8	0.003 6
Pitting	0.025 0	0.010 8	0.011 7	0.002 1	0.002 0
	0.035 1	0.011 2	0.012 4	0.002 9	0.002 9
	0.038 3	0.011 1	0.012 4	0.002 9	0.002 4
Uniform wear	1.881 7	0.238 1	0.421 3	0.064 9	0.082 8
	1.684 3	0.219 5	0.378 8	0.066 9	0.088 2
	2.060 5	0.269 8	0.439 6	0.061 5	0.083 4

Based on the basis images in Fig. 5 and the weight vectors in Tab. 2, we can interpret the relationship between the gearboxes' states and the corresponding bispectra as follows:

1) In any state, the basis image Fig. 5(a) far outweighs other basis images in constituting an original bispectrum, which means that the greatest amplitude of a bispectrum is always localized around the gear meshing frequency.

2) When pitting fault occurs, weights of basis images Figs. 5(b) and (c) increase. This phenomenon is in accord with the failure mechanism of local damage. The meshing process of a normal gearbox is composed of three stages, i. e. the meshing between a pair of teeth, then the meshing between two pairs of teeth, and finally the meshing between a pair of teeth. These two changes in stage generate two changes in the loads that act on gears, and two load changes result in two shocks in a vibration signal. When local damage occurs, load changes become more intense and the amplitude of the second harmonic of meshing frequency in a bispectrum increases simultaneously.

3) When uniform wear occurs, the weights of all basis images increase sharply. In particular, the weights of basis images Figs. 5(b) and(c) increase a lot more than those of other basis images. This phenomenon is in accord with the

failure mechanism of uniform wear. When uniform wear appears, involutes of teeth lose their shapes and the time-domain vibration signal develops gradually into the form of a square-wave. As a result, the amplitudes of meshing frequency and its higher-order harmonics increase in the bispectrum, and the higher the orders, the greater the increases[12].

4) The weights of basis images Figs. 5(d) and(e) in the three states indicate that amplitudes localize around frequency pairs which contain the difference and the sum of the meshing frequency and BPFO, accompanying each state of gearboxes.

The 45 training vectors and 15 testing vectors are then fed into an SVM classifier whose kernel is Gaussian with the bandwidth set at 10. Classification performance shows that the three different faults of testing samples are all correctly identified.

4 Conclusion

In this paper, a sparseness-controlled algorithm for NTF is proposed. The power of this algorithm is that it can lead to sparser basis images, less reconstruction error and less iteration steps than previous methods. Application of SNTF in extracting features from the bispectra of gearboxes' different states shows that this method results in sparse basis images whose energies are localized around fault characteristic frequencies, which help to explain the relationship between machinery faults and corresponding bispectra. It is also demonstrated that the proposed method is capable of detecting quadratic nonlinearity of the system, which is intangible in a bispectrum, and that the digital characteristics extracted from bispectra can contribute to effective fault classification. So, the proposed method alleviates the problems that bispectral analysis has in applications, and it is expected to play a positive role in the popularization of HOSA in machinery fault diagnosis.

References

[1] Yang Junyan, Zhang Youyun, Zhu Yongsheng. Intelligent fault diagnosis of rolling element bearing based on SVMs and fractal dimension [J]. *Mechanical Systems and Signal Processing*, 2007, **21**(5): 2012 − 2024.

[2] Zheng Haibo, Chen Xinzhao, Li Zhiyuan. Bispectrum based gear fault feature extraction and diagnosis [J]. *Journal of Vibration Engineering*, 2002, **15**(3): 354 − 358. (in Chinese)

[3] Huang Jinying, Bi Shihua, Pan Hongxia, et al. The research of higher-order cumulant spectrum for vibration signals of gearbox [C]//*Proceedings of IEEE International Conference on Information Acquisition*. Weihai, China, 2006: 1395 − 1399.

[4] Welling M, Weber M. Positive tensor factorization [J]. *Pattern Recognition Letters*, 2001, **22**(12): 1255 − 1261.

[5] Shashua A, Hazan T. Non-negative tensor factorization with applications to statistics and computer vision[C]//*Proceedings of the 22nd International Conference on Machine Learning*. Bonn, Germany, 2005: 793 − 800.

[6] Fitzgerald D, Cranitch M, Coyle E. Shifted 2D non-negative tensor factorization[C]//*Proceedings of IET Irish Signals and Systems Conference*. Dublin, Ireland, 2006: 509 − 513.

[7] Park S W, Savvides M. Estimating mixing factors simultaneously in multilinear tensor decomposition for robust face recognition and synthesis[C]//*Proceedings of Conference on Computer Vision and Pattern Recognition Workshop*. New York, 2006: 49 − 54.

[8] Cichocki A, Zdunek R, Choi S, et al. Non-negative tensor factorization using alpha and beta divergences [C]//*Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Honolulu, USA, 2007, **3**: 1393 − 1396.

[9] Zhang Qiang, Wang Han, Plemmons R J, et al. Tensor methods for hyperspectral data analysis: a space object material identification study [J]. *Journal of the Optical Society of America A: Optics and Image Science, and Vision*, 2008, **25**(12): 3001 − 3012.

[10] Hazan T, Polak S, Shashua A. Sparse image coding using a 3D non-negative tensor factorization [C]//*Proceedings of the 10th IEEE International Conference on Computer Vision*. Beijing, China, 2005: 50 − 57.

[11] Heiler M, Schnörr C. Controlling sparseness in non-negative tensor factorization[C]//*Proceedings of the 9th European Conference on Computer Vision*. Graz, Austria, 2006: 56 − 67.

[12] Ding Kang, Li Weihua, Zhu Xiaoyong. *Practical technology for fault diagnosis of gears and gearboxes* [M]. Beijing: China Machine Press, 2005: 91 − 109. (in Chinese)

含稀疏度约束的非负张量分解算法及其在故障诊断中的应用

彭 森 许飞云 贾民平 胡建中

(东南大学机械工程学院,南京 211189)

摘要:针对双谱分析在应用于机械设备故障诊断过程中面临的问题,提出了含有稀疏度约束的非负张量分解算法及基于此的二次故障特征提取方法. 首先,改进已有的非负张量分解算法,加入稀疏度控制策略;其次,将机械振动信号的双谱图像堆叠为一个三阶张量;然后利用改进后的分解算法对该张量进行二次故障特征提取,得到代表局部特征的"基图像";最后,通过计算得出基图像在构成原双谱图像中所占的权重,并将得到的权重向量用于故障分类. 将该方法应用于齿轮箱故障诊断的结果表明,从齿轮箱振动信号的双谱中提取出来的二次特征不仅能够反映出系统中存在的一些非线性特征,而且二次特征与故障特征频率之间有直观的对应关系,从而为解释齿轮箱故障与对应双谱之间的关系提供了很大的方便.

关键词:非负张量分解;稀疏度;特征提取;双谱;齿轮箱

中图分类号:TP206^{+}.3