WAVELET BASED FEATURE EXTRACTION OF BE STARS SPECTRA

Pavla Bromová
Doctoral Degree Programme (2), FIT BUT
E-mail: xbromo00@stud.fit.vutbr.cz

Supervised by: RNDr. Petr Škoda, CSc., Doc. Ing. Jaroslav Zendulka, CSc.
E-mail: skoda@sunstel.asu.cas.cz, zendulka@fit.vutbr.cz

Abstract:

The goal of our work is to create a feature extraction method for classification of Be stars spectra. This paper presents the feature extraction method based on wavelet transform and its power spectrum. The experimental verification of the feature extraction method is performed using clustering of Be stars spectra from the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic. We compare the results of clustering using different kinds of feature vector.

Keywords: Be star, spectrum, classification, clustering, feature extraction, discrete wavelet transform, wavelet power spectrum

1 INTRODUCTION

Technological progress and growing computing power are causing data avalanche in almost all sciences, including astronomy. The full exploitation of these massive distributed data sets clearly requires automated methods. One of the difficulties is the inherent size and dimensionality of the data. Efficient classification requires that we reduce the dimensionality of the data in a way that preserves as many of the physical correlations as possible.

Be stars are hot, rapidly rotating B-type stars with equatorial gaseous disk producing prominent emission lines in their spectrum [7]. Be stars show a number of different shapes of emission lines, as we can see in Fig.1. These variations reflect underlying physical properties of the star.

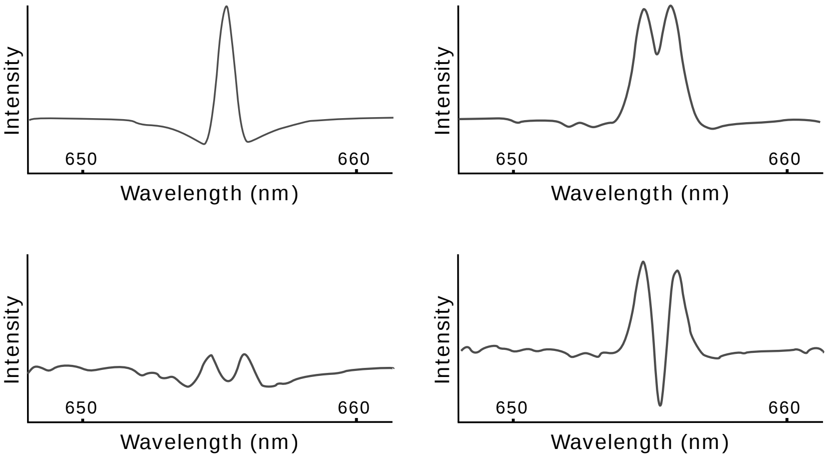

Figure 1: Typical shapes of emission lines in spectra of Be stars

Our work is focused on automated classification of Be stars based on a typical shape of their emission lines. There has not been much work on classification of Be stars. The only application found [1] is focused on a broader category of variable stars including pulsating Be stars. However, the method is not suitable for our goals, as it is applied on the whole spectrum where the local differences in the shapes of Be lines are lost. We need to zoom at the small part of a spectrum with the Be line and design a reduced, specific set of features characterizing and discriminating the shapes of Be lines. Due to a large variety of shapes, it is not easy to construct a simple criteria (like e.g. Gaussian fits) to identify the Be lines in an automatic manner.

In this paper, we present the feature extraction method based on wavelet transform and its power spectrum. The experimental verification of the feature extraction method is performed using clustering of Be stars spectra from the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic. We compare the results of clustering using different kinds of feature vector.

2 METHOD

The method is based on discrete wavelet transform and its power spectrum. A resulting feature vector is composed of two parts: 1. wavelet power spectrum, 2. value indicating the orientation of a spectral line (this information is lost in the wavelet power spectrum). The process of creating the feature vector is described here.

2.1 DISCRETE WAVELET TRANSFORM

The principle of discrete wavelet transform (DWT) can be described as passing the original signal through two complementary filters – low-pass and high-pass [5]. This results in two signals, referred to as approximation and detail. The approximation is a high-scale, low-frequency component of the signal, the detail is a low-scale, high-frequency component. After each pass through filters, downsampling (removing every alternative coefficient) is performed in order to avoid doubling the amount of data.

The decomposition process can be iterated by splitting the approximation part of a signal as it still contains some details. This can be repeated so long until we are satisfied with the resolution of components we have created. The wavelet transform of data at a level i of decomposition consists of approximation coefficients at i-th level and all detail coefficients up to i-th level, resulting in *number of levels + 1* coefficient bands. The wavelet coefficients reflect the correlation between the wavelet (at a certain scale) and the data array (at a particular location). A larger absolute value of a coefficient implies a higher correlation. For more details see e.g. [4].

In experiments up to now, we have used “symlet 4” wavelet and maximum possible level of decomposition = 5.

2.2 WAVELET POWER SPECTRUM

The wavelet power spectrum (WPS) is a useful way how to determine the distribution of energy within the signal [8]. By looking for regions of large power within WPS, we can determine which features of the signal are important.

The WPS at a particular decomposition level is calculated by summing up the squares of wavelet coefficients at that level [5]. For a set of wavelet coefficients $c_{j,k}$, where j is the level of decomposition and k is the order of the coefficient, WPS is given by:

$$wps(j) = \sum_{k=0}^{2^j-1} c_{j,k}^2$$

2.2.1 RECTIFICATION

It has been shown in [3] that wavelet power spectra are biased in favor of low frequencies. For example, for a signal comprising two sine waves of the same amplitude but distinct frequencies, a wavelet analysis will yield two spectral peaks of different magnitude, the one on the low frequency being larger. This counters our expectation and is also in contrast to the result of any classical global analysis (such as Fourier transform), making comparison of the peaks across the scales impossible.

In [3], they established theoretically that the bias actually results from the traditional definition of "energy" for the wavelet power spectra which is not physically consistent. They present a physically consistent definition of energy: the transform coefficients squared divided by the scale they associate. The traditional biased power spectra are therefore easily rectified.

2.3 NORMALIZATION

WPS is normalized so that its total energy equals to 1, so it consists of percentages of energy corresponding to individual levels.

2.4 ORIENTATION OF SPECTRAL LINE

The information about the orientation of a spectral line is lost in the wavelet power spectrum, so we need to add it somehow into the feature vector. We want to distinguish whether a spectral line is oriented up (emission line) or down (absorption line), so we use one positive and one negative value. The question is which absolute value to choose. So far we have tried three values: 1, 0.1, and the amplitude of a spectral line, measured from the continuum of value 1.

3 EXPERIMENTS

The experimental verification of the feature extraction method is performed using clustering. So far, the whole process has been implemented in Matlab, using its embedded algorithms. The stages are described in following sections.

3.1 DATA SELECTION

We use spectra of Be stars from the archive of the Astronomical Institute of the Academy of Sciences of the Czech Republic. The collection contains around 10000 samples from approximately 700 stars. We select one sample from each star, so our data set contains approximately 700 samples.

We also select only a part of a spectrum which contains the emission line, so that the sample has 256 values and the emission line is in the center. The number 256 was chosen according to average width of emission line and according to DWT requirements.

Most of the spectra are normalized (lying on a continuum of a value 1), but a few of them are not, which is caused by a mistake of a normalization procedure. We filter these unnormalized samples by selecting only those samples which contain at least one value less than 2.

3.2 FEATURE EXTRACTION

Feature extraction method is described in previous chapter. From resulting coefficients we create different kinds of feature vector:

- **Spectrum**: original spectrum values, normalized to range [0,1]. (In this case coefficients are not used.)
- **Approximation**: approximation coefficients, normalized to range [0,1].

- **Approximation + detail**: approximation and detail coefficients of last level, normalized to range [0,1].
- **N largest**: 10 largest absolute values of coefficients, normalized to range [-1,1].
- **WPS + orientation 1**: one part of a feature vector is wavelet power spectrum, normalized so that its total energy equals to 1. Second part of a feature vector is a value indicating the orientation of a spectral line – lines oriented up have the value 1, lines oriented down have the value -1.
- **WPS + orientation 0.1**: the same as the previous one, except the absolute value of orientation 0.1.
- **WPS + amplitude**: one part of a feature vector is normalized wavelet power spectrum as in the previous case. The second part is the amplitude of emission line measured from the continuum of value 1.

3.3 Clustering

So far, k-means algorithm was used for clustering. In k-means, a number of clusters has to be determined, which is also one of the goals of the experiments. We have tested 3-8 clusters, based on the results of manual sorting of spectra performed by experts. Clustering was performed in 30 iterations and results were averaged.

3.4 Evaluation

The silhouette method [6] is used for the evaluation of clustering and comparison of different kinds of feature vector. The silhouette value for each object is a measure of how similar that object is to objects in its own cluster compared to objects in other clusters, and ranges from -1 to $+1$, where $+1$ is the best result.

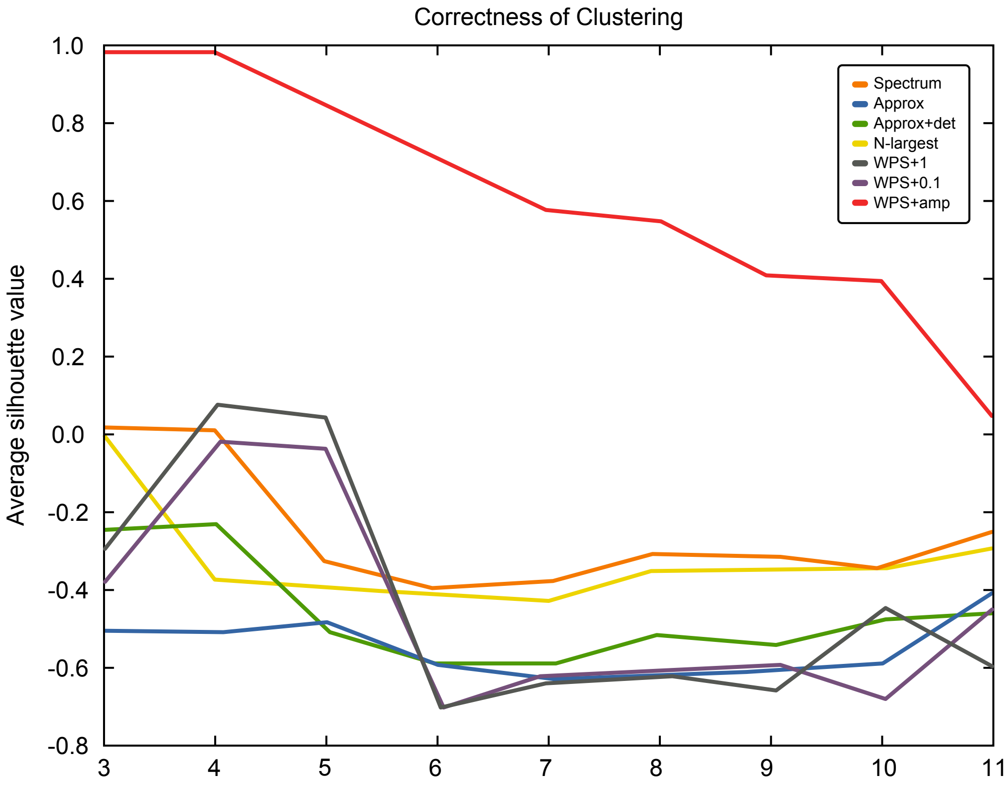

Figure 2: Correctness of clustering using different feature vectors

As we don't have any reference method for Be stars for comparison, we compare our results with a common way of feature extraction from time series using wavelets – keeping N largest coefficients of wavelet transform [2].

4 RESULTS

Fig. 2 shows the average silhouette value for different kinds of feature vector and different number of clusters. The best results are given by the feature vector consisting of WPS and the amplitude of emission line, which indicates that the amplitude is the most discriminating feature of spectra. The other two kinds of feature vector containing WPS perform quite well considering their smaller length in comparison with other feature vectors.

5 CONCLUSION

In this paper, we have analysed the capabilities of using wavelet power spectrum for classification of spectra of Be stars. We have proposed a feature extraction method based on discrete wavelet transform and its power spectrum. We have created different kinds of feature vector and compared them on clustering. The results show that feature vectors containing WPS perform quite well considering their smaller length in comparison with other feature vectors. They also indicate that the amplitude is the most discriminating feature of spectra, which requires further analysis of usability of this feature.

In future, the continuous wavelet transform is assumed to be tested, based on the recent research of literature which indicates that it should be more suitable for our purpose. More experiments with different types of wavelet and another kinds of feature vector are also assumed.

ACKNOWLEDGEMENT

This work was partially supported by the research plan MSM0021630528, the specific research grant FIT-S-11-2 and the IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

REFERENCES

[1] J. Debosscher. *Automated Classification of variable stars: Application to the OGLE and CoRoT databases*. PhD thesis, Institute of Astronomy, Faculty of Sciences, Catholic University of Leuven, 2009.

[2] T. Li, S. Ma, and M. Ogihara. Wavelet methods in data mining. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 553–571. Springer, 2010.

[3] Y. Liu, X. San Liang, and R. H. Weisberg. Rectification of the bias in the wavelet power spectrum. *Journal of Atmospheric and Oceanic Technology*, 24(12):2093–2102, 2007.

[4] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.

[5] S. Prabakaran, R. Sahu, and S. Verma. Feature selection using haar wavelet power spectrum. *BMC Bioinformatics*, 7:432, 2006.

[6] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53 – 65, 1987.

[7] O. Thizy. Classical Be Stars High Resolution Spectroscopy. *Society for Astronomical Sciences Annual Symposium*, 27:49, 2008.

[8] C. Torrence and G. P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79:61–78, 1998.