Auto-flag the baseline for Mingantu Ultrawide Spectral Radioheliograph with LSTM

Jun Cheng, Long Xu, Xuexin Yu, Linjie Chen, Wei Wang, Yihua Yan

Key Laboratory of Solar Activity, National Astronomical Observatories,
Chinese Academy of Sciences, Beijing, China

chengjun@nao.cas.cn, lxu@nao.cas.cn, yuxuexin2008@gmail.com, ljchen@nao.cas.cn,
wwang@nao.cas.cn, yyh@nao.cas.cn

Abstract— **Mingantu Ultrawide Spectral Radioheliograph (MUSER) is an aperture synthesis telescope consisting of a group of small antennas to image the Sun. Each two antennas form a baseline contributing a Fourier sampling point for each time of imaging. Both amplitude and phase of a baseline form a time sequence in a time interval. Normally, amplitude/phase should vary linearly in a short time interval. However, many reasons would result in errors of baselines, damaging image quality. This work makes the first attempt to auto-flag baselines so as to delete bad baselines and improve image quality. Inspired by the big success of long short-term memory (LSTM) for time series analysis, we believe that LSTM could accomplish auto-flagging (a binary classification task) of baselines even better by treating them as time sequences. Thus, LSTM is employed to learn the representation of the phase of each baseline for classification, where the interaction and connection within a phase sequence are explored, benefiting classification. The experimental results demonstrate that LSTM can well capture the characteristics of the phase of a baseline, and thus achieves better classification.**

Index Terms—**Auto-flag; MUSER; LSTM; Phase; Classification**

I. INTRODUCTION

The solar activities such as coronal mass ejections (CMEs), flares, and solar energetic particles (SEPs), etc., have great influence in the space weather studies [1]. Radio observations provide important diagnosing tool on the related parameters of solar activities [2], such as the magnetic field, electron density, plasma temperature, etc. MUSER is a new solar-dedicated radio heliograph employs the Aperture Synthesis principle, which has the capability to image the Sun with plentiful solar radio activities at multiple frequencies (0.40-15.00GHz) with high time resolution, high spatial resolution and high frequency resolution [3]. 100 antennas are arranged into two arrays, MUSER-I and MUSER-II. MUSER-I consists of 40 antennas of 2.5m diameter, and MUSER-II consists of 60 antennas of 2m diameter. Each two antennas in an array form an interferometer, which records a Fourier component of a sky object each time. For MUSER-I, there are 780 (=40×39/2) interferometers totally, so 780 Fourier components are recorded each time. These Fourier components of a sky object are named as visibility, while its spatial image is named as brightness. Through inverse Fourier transform, we can convert a visibility into a brightness, i.e., a spatial image of the sky object [4].

We also name the system of each two antennas a baseline in an array of Aperture Synthesis. During the observation, some baselines may break down due to a variety of reasons, e.g. instrumental failure, radio frequency interference, ionospheric scintillations etc. [5]. The recorded data of these baselines would pollute the rest of data, so it is necessary to figure out these baselines and excluded them from the others for better imaging of the following step. Recognition of the distorted or abnormal data, so called data flagging, is an important step in data processing of radio astronomy observation, which would have a significant impact on the final imaging.

In the early days of radio research, data flagging is generally done manually. With the new radio telescopes put into use, such as Japan's Nobeyama Radioheliograph [6], the France's Nancay Radio Heliograph [7], the Russian's Siberian Solar Radio Telescope [8], China's MUSER and so on, the amount of data accumulated by these telescopes has risen from the GB level to the TB level. Especially for MUSER, which consists of a high frequency array with 60 antennas and 528 channels, and a low frequency array with 40 antennas and 64 channels, produces more than 3TB data per day, the manual processing has been completely impractical. Besides, there is some uncertain during artificial markings. Therefore, auto-flag has great significance for solar radio astronomy study.

A diverse range of approaches have been taken for identifying interfering signals in radio astronomical data (see e.g. [9–13]. We note that general purpose packages like Astronomical Image Processing System (AIPS) and Common Astronomy Software Applications (CASA) also have tools that flag data. As for MUSER, in 2015, Meng Cheng [14] tried to use FLAGCAL like the methods used by Giant Metrewave Radio Telescope (GMRT), but the results were unsatisfactory. In 2016, DAI Hui-mei [15] used the Support Vector Machine (SVM) technology, the results show that the SVM is a robust approach to flag the MUSER visibility data, and could reach the accuracy of about 86%. Nowadays, with the availability of massive data, deep learning [16] has been extensively explored to solve many traditional tasks, such as recognition, classification, regression and clustering, with state-of-the-art performances. The most important is this method can directly learn useful features from the unlabeled or labeled data, instead of the need for hand-engineering as SVM.

Long short-term memory (LSTM) is one of the common network used in deep learning, which tries to explore the relations and interactions of sequential data. Thus, LSTM is very suitable for representation learning of data related to time. As mentioned above, a baseline gives a Fourier component represented by a complex number each time, so the visibility phase/magnitude of a baseline forms a time sequence in a certain time interval. According our observation as explained in Section II, the visibility phase of a baseline can be referred to judge the status of a baseline. Thus, LSTM is employed to learn the representation of visibility phase and further classify baselines into bad and good ones in this work.

The rest of this paper is organized as follows. Section II presents the details of the auto-flagging model by using LSTM. Section III provides experimental results on phase classification. The conclusions are given in the last section.

II. PROPOSED MODEL OF AUTO-FLAGGING

For tuning MUSER to a good condition, several techniques, such as delay compensation, fringe stop are conducted in MUSER data processing. For a target of a point source like satellite, its visibility phase of each baseline should be stable in the time frequency plane. Otherwise, there is error in the recorded data. Thus, we can judge whether a baseline is in a good condition by its phase.

A. Pre-processing of visibility phase

A data file is recorded per one minute for solar radio observation of MUSER-I. The time resolution of MUSER-I is 3ms, so one frame is acquired every 3ms, the size of each frame is 100,000 bytes. Fig. 1 shows the details of the data format of MUSER-I, where the cross-correlation coefficients are preceded by file header and some reserved bytes. The cross-correlation is recorded by a baseline, while auto-correlation is given by an antenna itself. To judge a baseline, we use cross-correlation data.

MUSER-I operates on multiple frequency channels. Here, only one frequency channel is randomly selected for analyzing. Taking the 53-th channel for example, the corresponding frequency is 1.7 GHz. Since the number of antennas of MUSER-I is 40, there would be 780 baselines, and thereby 780 cross-correlation complexes per 3ms. The data of 5 minutes is collected to give the samples of visibility phases of baselines. At the same time, the original 2400 frames are down-sampled into 240 frames evenly per one minute to reduce the complexity of network training. Thus, the phase matrix of all baselines is of the size 1200×780 in a time interval of 5 minutes. For each baseline, the phase is a vector of the length 1200.

Given a complex, its phase is calculated by

$$\varphi = \arctan\left(I_{real}/I_{image}\right) \tag{1}$$

where I_{real} and I_{image} are the real part and imag part of cross-correlation respectively.

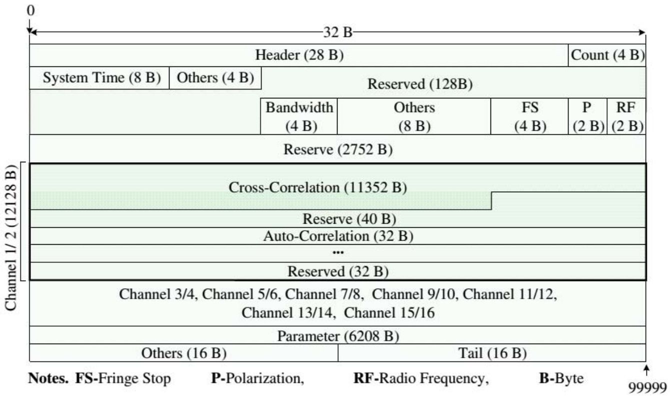

Fig. 1 Data format of MUSER- I

B. Phase plotting and flag label

For establishing auto-flagging system by machine learning, a large amount of labelled data is needed. We firstly establish a phase database including the data of 11 days. Each day contains 5 minutes of data, given by a 1200×780 matrix. Thus, the whole database can be given by a 1200×780×11 matrix and the corresponding 1×780×11 labels. The labels are given by directly analysing the image of each baseline at a time interval of 5 minutes as shown in Fig. 2. Only two types of labels: "good" and "bad" (1 and 0) are assigned. Simply, if the data distributes randomly along time, like A0-B2, label of "bad" is assigned. This situation indicates the phase extremely unstable, so the corresponding baseline is out of service. If the curve is almost straight or partially straight or even with some sharps, as shown in A0-A6, A0-C11, A1-C7 in Fig. 2, the baseline work well, so it is marked as "good". The details of the database are list in TABLE I. The training and testing are performed on the split of the database. 6016 "good" and 1784 "bad" are randomly selected from the dataset for training, and the rest are for testing. The details of training set and testing set are listed in TABLE II.

TABLE I
DETAILS OF THE DATABASE

Phase type	Good	Bad	Total
Phase number	6649	1931	8580
Phase size	1×1200	1×1200	1×1200

TABLE II
DETAILS OF TRAINING AND TESTING SPLIT

Phase type	Good	Bad	Total
Training Number	6016	1784	7800
Testing Number	633	147	780
Phase Size	1×1200	1×1200	1×1200

978-1-5386-0462-5/17/$31.00 ©2017 IEEE. VCIP 2017, Dec. 10 – 13, 2017, St Petersburg, U.S.A.