Separation of magnetotelluric signals based on refined composite multiscale dispersion entropy and orthogonal matching pursuit

Xian Zhang^{1,2,3,4}, Jin Li^{2,3,*}, Diqian Li^{1,4,*}, Yong Li³, Bei Liu⁵, Yanfang Hu^{1,4}

¹ Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment, Monitoring Ministry of Education, School of Geoscience and Info-Physics, Central South University, Changsha 410083, China.

² Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China.

³ Key Laboratory of Geophysical Electromagnetic Probing Technologies of Ministry of Natural Resources, Institute of Geophysical and Geochemical Exploration, Chinese Academy of Geological Science, Langfang 065000, China.

⁴ Hunan Key Laboratory of Nonferrous Resources and Geological Hazard Exploration, Changsha 410083, China.

⁵ College of Mathematics and Physics, Hunan University Arts and Science, Changde 415000, China.

*Correspondence: geologylj@163.com; lidiquan@csu.edu.cn

Abstract: Magnetotelluric (MT) signal processing can increase the reliability of data. Traditional MT de-noising methods are usually filtered in entire MT time-series sequence, which result in losing of useful MT signals and the degradation of electromagnetic inversion imaging accuracy. However, targeted MT noise separation will retain the part of data that is not affected by strong noise, and enhance the quality of MT data. Thus, we proposed a novel method which using refined composite multiscale dispersion entropy (RCMDE) and orthogonal matching pursuit (OMP) for MT noise separation. Firstly, we extracted the RCMDE characteristic parameters from each segment of the MT time-series. Then, the characteristic parameters are input to the fuzzy c-mean (FCM) clustering for automatic signal-noise identification. Next, the identified noise segments were independently denoised by using OMP method. Finally, the reconstructed signal consists of the denoised data segments and the identified useful signal segments. We conducted the simulation experiments and algorithm evaluation on the EMTF data, simulated data and measured sites. The results indicate that a variety of entropy without scale factor is not stable enough, and the RCMDE can improve the stability of multiscale dispersion entropy (MDE) and multiscale entropy (MSE) by analyzing the characteristics of the signal samples library, and thus effectively dividing MT signals and noise. In this paper, the existing techniques of the entire time domain de-noising and signal-noise identification method are compared. Only for RCMDE and OMP as characteristic parameter and noise separation method, the proposed method simplified the multi-features fusion method, and improved the accuracy of signal-noise identification, accelerated de-noising efficiency, and improved the MT data quality of low-frequency band.

Keywords: magnetotelluric (MT); refined composite multiscale dispersion entropy (RCMDE); orthogonal matching pursuit (OMP); noise separation

Introduction

The magnetotelluric (MT) sounding method is one of the most mature electrical exploration techniques in recent years (Tikhonov 1950; Cagniard 1953), which is mainly used to measure the

orthogonal electric and magnetic field at the Earth's surface of the geoelectrical structures, electrical exploration of minerals, and electromagnetic fracture monitoring (Becher and sharpe 1969; Vallianatos 1996; Qi et al. 2020). Due to the wide frequency range of natural MT signal, it is vulnerable to interfere from an increasingly wide range of natural and human environmental noise. Therefore, effectively suppressing noise, improving the signal-noise ratio (SNR), and ensuring the quality of MT data have always been the core issues of MT data collection and data processing. However, the characteristics of MT signals are non-linear, non-stationary and non-minimum phase, which does not meet the conditions of Fourier transform (Hermance 1973). Therefore, strong interference of MT data will cause the excessive distortion of the apparent resistivity-phase curve and excessive concentration of phase angle in polarization direction.

In order to suppress the MT strong interference and improve the quality of MT data, numerous de-noising techniques have been applied. The remote reference method (Gamble et al. 1978; Gamble et al. 1979) and the robust impedance estimation method (Egbert and booker 1986; Jones et al. 1984) are the most commonly MT data processing method. The remote reference method relies on the selection of reference sites, while the robust impedance estimation method frequently fails to the effect of correlated noise and persistent strong interference. Some novel data processing methods are applied to the field of MT noise suppression. For example, the wavelet transform can effectively suppress the local electromagnetic related noise (Trad and travassos 2000), while it relies on the selection of the wavelet basis function. With the increase of scales, the spectral localization of the corresponding orthogonal basis function deteriorates, limiting the fine decomposition of MT signals. The Hilbert-Huang transform (Huang et al. 1998) has been applied in the electrical method exploration, which can effectively suppress the MT signal interfered by power frequency (Cai et al. 2009). Compared with wavelet transform, it has too stronger time-frequency characterization capabilities to choose the basis function. Mathematical morphological filtering can effectively suppress large-scale interference and baseline offset in the MT signal, and maintain the local characteristics of the target signal (Tang et al. 2012), but it difficult to select the types and sizes of structural elements. By analyzing the strong electromagnetic interference of the MT data, it mainly comes from human facilities such as power lines, engineering machinery and equipment, and has certain regular. Wang et al. proposed inter-station transfer functions and neighboring site method to recover natural electric and magnetic fields, and reduce the influence of anthropogenic noise (Wang et al. 2017). Variational mode decomposition (VMD) is a novel modal decomposition algorithm, which has been applied for MT noise suppression. Since the K value in VMD algorithm can only be manually selected. Li J et al. combined VMD with detrended fluctuation analysis (DFA) to adaptively select K value and the number of the reconstructed modals, thus improving the de-noising effect (Li et al. 2020). In addition, the original MT signal-noise identification method is based on fractal, entropy, and complexity, and multi-feature cluster analysis, and the fidelity of the data is improved (Li et al. 2018). By removing noises in time domain, the statistical analysis and time-series editing methods can directly and effectively improve the quality of MT data, but they could also damage the effective slow portion on the noisy segments. Although we extract multiple description MT signal-noise characteristics for identifying MT strong interference, the time and efficiency are spent in computing features. In order to overcome the above-mentioned issue, we introduce a single feature to describe the difference between signal and noise on multiple scales.

Entropy shows the degree of chaos. It was first used by physics to define the degree of material chaos, and it later extended to thermodynamics and informatics to describe the dynamic process of a certain event becoming chaotic. Entropy is widely used now, such as sample entropy, fuzzy entropy and approximate entropy, which can be used to measure the complexity of the system. Based on multi-scale analysis on the entropy theory. Multiscale entropy (MSE) was proposed to quantify the complexity of signals on multiple temporal scales. Multiscale dispersion entropy (MDE) is a parameter to evaluate the multiple scales dynamic complexity of time-series (Zhang et al. 2018). The refined composite MDE (RCMDE) can increase the accuracy of entropy estimation and decrease the probability of the faced with the undefined entropy, and overcome the shortcoming of multiscale entropy (MSE) and MDE, which can combine the information of multiple coarse-grained sequences, reduce the standard deviation of entropy, make the entropy

more reliable, and improve numerical stability (Azami et al. 2017). The FCM clustering algorithm is an unsupervised method for data analysis and modeling, and it widely used in data classification and pattern recognition (Xu et al. 2019; Zhang et al. 2019), which uses the input features to generate the clustering center, calculates the Euclidean distance between the clustering points and the clustering centers, and obtains the membership of the clustering centers to divide the types of input features automatically. The sparse representation is to represent the signal with as few atoms as possible in a given over-complete dictionary, so as to obtain a more concise representation of the signal and acquire the contained information and process signal more conveniently. Based on MP algorithm, the orthogonal matching pursuit (OMP) algorithm is also a classic greedy algorithm (Pati et al. 1993; Mallat and Zhang 1993). By using the Gram-Schmidt orthogonalization method to orthogonal the selected optimal atom and atomic set, so that the selected atom will be orthogonalized to the residual in each iteration, thereby accelerating the convergence speed of the algorithm. In view of the fact that the MT signals usually contain the stationary and non-stationary components, an atomic library composed of the sine, cosine atoms and wavelet atoms are designed to realize the adaptive and accurate matching of MT signals.

In the paper, we proposed a novel method using RCMDE and OMP to separate MT noise. The RCMDE characteristic is associated with the fuzzy c-means (FCM) clustering for MT signal-noise identification, and using orthogonal matching pursuit (OMP) algorithm for MT noise separation. We first apply the distribution of various entropies in the sample library, further verified the stability of RCMDE and the de-noising effect of OMP, and further applied to EMTF simulation and measured signal analysis. The proposed method is compared with the remote reference method and the OMP-based overall method, which can purposefully remove the identified noise and retain the low-frequency useful MT signal. Compared with the fractal-entropy and clustering method (Li et al. 2018) in the analysis of multiple characteristics of MT data, it makes more precise and efficient features extraction of MT data that improves the efficiency of MT signal-noise identification. Moreover, the experimental results at the measured sites of apparent resistivity-phase curves and polarization direction reflect that the proposed method can more reliably approach to the natural MT field, and truly reflect the underground electrical structural information.

Methods and Materials

In view of the long acquisition time of MT sounding data, the degree of interference by noise is also not completely consistent. We search for the robust characteristic parameters to quantitatively identify the data segments that are not interfered by noise or less interfered, and retain the MT useful signal from being over-processed, only by eliminating the abnormal waveforms can the accuracy of MT noise suppression be improved, and obtain the high-resolution MT signals. Thus, the proposed method compares with the MSE and MDE, and extracts RCMDE feature from the time-series waveforms and inputs to FCM cluster analysis, then using OMP algorithm for noise separation. Based on the above mentioned, the following respectively introduces RCMDE as unique feature extraction and OMP as de-noising algorithm.

Dispersion Entropy (DE)

Dispersion entropy (DE) is proposed by Rostaghi M and Azami H in 2016, which is a nonlinear dynamics method to characterize the complexity and irregularity of the time series (Rostaghi and Azami 2016; Mitiche et al. 2018).

- (1) Suppose $x_j = j = 1, 2, \dots, N$ is mapped to c classes with integer indices from 1 to c .

To realize the purpose, the normal cumulative distribution function (NCDF) maps x into $y = y_1, y_2, \dots, y_N$ from 0 to 1 as follows:

$$y_j = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x_j} e^{-\frac{t-\mu}{2\sigma^2}} dt \quad (1)$$

where σ and μ are the standard deviation (SD) and mean of time series x , respectively. Then, a linear algorithm is employed to each y_i to an integer which ranges from 1 to c . For each member of the mapped signal, we use $z_j^c = \text{round } c \cdot y_j + 0.5$, where z_j^c denotes the j^{th} member of the classified time series and rounding involves either increasing or decreasing a number to the next digit (Rostaghi and Azami 2016). Note that, although this part is linear, the entire mapping method is non-linear due to the use of the NCDF.

- (2) Each embedding vector $z_i^{m,c}$ are made with embedding dimension m and time delay d , according to construct the time series $z_i^{m,c} = z_i^c, z_{i+d}^c, \dots, z_{i+m-1}^c, i=1, 2, \dots, N-m-1$ (Rostaghi and Azami 2016; Bandt and Pompe 2002) which is mapped to a dispersion pattern $\pi_{v_0 v_1 \dots v_{m-1}}$, where $z_i^c = v_0, z_{i+d}^c = v_1, \dots, z_{i+m-1}^c = v_{m-1}$. The number of possible dispersion patterns that can be assigned to each time series $z_i^{m,c}$ is equal to c^m , since the signal has m members and each member can be one of the integers from 1 to c (Rostaghi and Azami 2016).

- (3) For each of c^m potential dispersion patterns, relative frequency is obtained as follows:

$$p_{\pi_{v_0 v_1 \dots v_{m-1}}} = \frac{\text{Number } i | i \leq N-m-1, z_i^{m,c} \text{ has type } \pi_{v_0 v_1 \dots v_{m-1}}}{N-m-1} \quad (2)$$

where $p_{\pi_{v_0 v_1 \dots v_{m-1}}}$ denotes the number of dispersion patterns $\pi_{v_0 v_1 \dots v_{m-1}}$ that are assigned to $z_i^{m,c}$, divided by the total number of embedded signals with embedding dimension m .

- (4) The DE value is derived from the definition of Shannon's entropy and is defined as follows:

$$DE_{x,m,c,d} = - \sum_{\pi=1}^{c^m} p_{\pi_{v_0 v_1 \dots v_{m-1}}} \cdot \ln p_{\pi_{v_0 v_1 \dots v_{m-1}}} \quad (3)$$

According to the differences of various entropies in the sample library signal which include the 150 data samples that are from MT field measurements, refer as (Li et al. 2018). Among them, 50 data samples without interference are from the unmanned areas, and the remaining 100 data samples are collected from the strong electromagnetic interference areas. Figure 1 shows the results of extracting various entropies from sample library signals, such as sample entropy (SE), fuzzy entropy (FE), approximate entropy (AE) and dispersion entropy (DE).

Fig. 1 The results of (a)SE, (b)FE, (c)AE and (d)DE using sample library signals.

It can be seen from Figure 1 that using different entropies for the sample library can represent different entropy values, and can distinguish MT data without interference and with interference. While the discrimination of AE is the smallest, and SE and FE have larger fluctuation in the MT data without interference and with interference. Taking into account the signal amplitude and is less affected by the mutation signal, the DE value of the MT data without interference is more stable. MT data without interference is also what we need to identify and retain in the follow-up. However, the entropy with a single scale cannot fully extract the complexity characteristics of the MT signal-noise. Next, a refined composite technique is used to improve the DE on the multi-scale.

Refined Composite Multiscale Dispersion Entropy (RCMDE)

Multiscale dispersion entropy (MDE) is the combination of the coarse-graining (Costa et al. 2005) with DE, and then calculate the DE value of the coarse-graining sequence to obtain the DE at different scale. Instead, the mapping based on the normal cumulative distribution function (NCDF) used in the calculation of DE for the first temporal scale is maintained across all scales. And refined composite MDE is improved MDE algorithm as follows:

For scale factor τ which has different time series, corresponding to different starting points of the coarse-grained process are created, and the RCMDE value is defined as the Shannon entropy value of the average of the dispersion patterns of those shifted sequences (Azami et al. 2017). The k^{th} coarse-grained time series $x_k^\tau = x_{k,1}^\tau, x_{k,2}^\tau, \dots$ of u is as follows:

$$x_{k,j}^\tau = \frac{1}{\tau} \sum_{b=k+\tau}^{k+\tau j-1} u_b, 1 \leq j \leq N, 1 \leq k \leq \tau \quad (4)$$

Then, for each scale factor, RCMDE is defined as follows:

$$RCMDE(x, m, c, d, \tau) = - \sum_{\pi=1}^{c^m} \bar{p}_{\pi_{v_0 v_1 \dots v_{m-1}}} \cdot \ln \bar{p}_{\pi_{v_0 v_1 \dots v_{m-1}}} \quad (5)$$

where $\bar{p}_{\pi_{v_0 v_1 \dots v_{m-1}}} = \frac{1}{\tau} \sum_{k=1}^{\tau} p_k^\tau$ with the relative frequency of the dispersion pattern π in the time series $x_k^\tau, 1 \leq k \leq \tau$.

Thus, we predefined the parameter values, refer to (Rostaghi and Azami 2016), such as embedding dimension m is 2, number of classes c is 6, time lag d is 1, scale factor τ , that is corresponding to τ sets of characteristic parameters in this paper. Figure 2 shows the results of MSE, MDE and RCMDE using a set of sample library signals with different scale factor. The

sample library signals are used for feature extraction and FCM clustering analysis to distinguish signal and noise.

Fig.2 The results obtained for (a)MSE, (b)MDE and (c)RCMDE using a set of sample library signals at different scale factor, among them, the abscissa represents the scale factor τ , and the ordinate represents the entropy value at the corresponding scale.

We extract the characteristic parameters using only MT data without interference and MT data with interference (square wave interference and triangle wave interference). From Figure 2, a set of sample library signals have different entropy values at different scales. When the scale factor is increased, the difference between MSE and MDE of a set of sample library signals is low, crossover phenomenon and higher oscillations of the different interference curves, which is not conducive to classify noise and signal. However, the RCMDE shows the importance of the refined composite technique to improve the stability of results. Moreover, extracting appropriate characteristic parameters to describe the features of the MT data is helpful to improve the clustering effect. As can be seen from Figure 2(c), when the scale factor is 1, the RCMDE feature value is the largest, and the single characteristic value will be meaningless, which cannot reflect the scale characteristic of the MT data. With the scale increases, these curves become more stable, the size of the scale will determine the number of characteristic parameters. The larger the scale factor is, the more characteristic parameters values will be generated, and the accuracy and efficiency of signal-noise identification of the massive measured data also will decline. Reference (Li et al. 2018) showed that the four characteristic parameters are extracted from different angles, FCM clustering analysis as signal-noise identification, and the strong interference is also identified. When the signal and noise are slightly blurred, these different characteristic parameters will be difficult to divide the signal and noise. In the paper, the RCMDE can generate multiple feature parameter values to represent MT signal-noise at different scale factor, while many characteristic parameters will reduce the clustering

accuracy and consume more time. Thus, we only select the RCMDE scale factor of 2 as the feature vector analysis.

De-noising Method

Traditional matching pursuit algorithm at each iteration only ensures that matching residual data is orthogonal with an atom, which is prone to local optimization and result in a low matching accuracy and a large amount of computation (Jin et al. 2014). The orthogonal matching pursuit (OMP) algorithm is based on matching pursuit (MP) algorithm, and also an improved version of MP by ensuring full backward orthogonality between the matching residual and the selected waveforms at each iteration, which ensures the optimal approximation with respect to all the selected subset of the dictionary after any finite number of iterations (Wang et al. 2013).

Given a signal x , that length is N . $D = [g_\gamma]_{\gamma \in \Gamma}$ is an over-complete dictionary, that is the Fourier atomic library and wavelet atomic library, which for signal sparse decomposition. g_γ is the γ_{th} atom in the dictionary set Γ , and $\|g_\gamma\| = 1$.

Initialized signal residual $R^0 = x$, and reconstructed signal $\bar{x}_m = 0$, we selected the atom set Ψ_0 is empty set, the number of iterations $m = 1$, the maximum number of iterations M . When the iteration stop condition is not met, repeat the following steps (Tropp and Gilbert 2007).

- (1) We select the atom g_γ that most closely matches the analysis signal x from the dictionary. They meet the following conditions:

$$\left| \langle R^{n-1}, g_\gamma^m \rangle \right| = \sup_{\gamma \in \Gamma} \left| \langle R^{n-1}, g_\gamma^m \rangle \right| \quad (6)$$

- (2) Update the selected atom set $\Psi_m = \Psi_{m-1} \cup g_\gamma^m$.
- (3) Find the projection coefficient according to the least square method $u_m = \Psi_m^T \Psi_m^{-1} \cdot \Psi_m^T x$. Thus, the reconstructed signal is $\bar{x}_m = \Psi_m x_m$, and residual signal is $R^m = x - \bar{x}$.
- (4) Update the number of iterations $m = m + 1$. Judge whether the energy ratio $\|R^m\|_2 / \|x\|_2$ of the residual signal to the original signal is less than the given value. If not satisfied, return to step (1). If satisfied, the reconstructed signal is obtained $x = \bar{x}_m$ and residual signal is $R = R^m$.

Figure 3 shows the de-noising effect of MP and OMP methods for the noisy signal with 20db White Gaussian Noise, and the length of the noisy signal is 2048. Among them, the mentioned atomic library consists of sine (sin) and discrete cosine (dct) atoms, and the symlets (sym) and daubechies (db) wavelet atoms.

Fig. 3 The de-noising effect and frequency spectrum analysis of the noisy signal with (a) matching pursuit (MP) and (b) orthogonal matching pursuit (OMP).

To estimate the de-noising effects of MP and OMP methods for the above noisy signal, the normalized cross-correlation (NCC), signal-noise ratio (SNR), mean square error (MSE) and the time factor (T) were used for the quantitative analysis. Definitions of these parameters are as refer to (Li et al. 2020).

Table 1 is the comparison between the de-noising performance of MP and OMP methods.

Table 1 The de-noising performance of MP and OMP methods.

Methods	NCC	SNR(dB)	MSE	T(s)
MP	0.9848	15.1976	0.0651	2.45
OMP	0.9989	26.3956	0.0049	0.59

As can be seen from Figure 3 and Table 1, by comparing the de-noising effects of MP and OMP methods with the same number of atoms and iterations, we found that the MP method still has

residual noise in the reconstructed signal, while the OMP method shows the excellent characteristics in the NCC, SNR, MSE and Time, and its frequency spectrum also strictly corresponds to the left signal. Therefore, OMP method ensures that the residue is orthogonal to the all selected atoms, and converge speed is faster than MP method. The OMP method is more suitable for subsequent noise separation, and also requires much less memory and computation time.

Experiments

Clustering Analysis of the Sample Library

In this section, we applied the FCM clustering analysis for sample library signals through feature extraction based on RCMDE when the scale factor is 2. The FCM clustering algorithm obtains the membership degree of each sample to all clustering centers by optimizing the objective function, thereby determining the type of the samples to achieve the purpose of automatic classification.

Figure 4 is the FCM clustering effect of library signals and type classification.

Fig. 4 (a) is FCM clustering effect of sample library signals and (b) is type classification. Among them, characteristic X and Y in the coordinate axis represent the refined composite multiscale dispersion entropy (RCMDE) value when the scale factor is 1 and 2 in the left figure, respectively.

From Figure 4 shows that the FCM clustering algorithm effectively divides MT data without interference and with interference, among them, two pink circles represent two different types. According to verify with the FCM clustering effect, as shown in the right figure, MT data without interference are classified into the Type 2, and MT data with interference are classified into Type 1 which can be purposefully removed.

Numerical Simulation Analysis of the EMTF Data

The numerical simulation analysis is based on the open source software EMTF, which is an open source code for the MT single-site robust estimation and remote reference analysis written by Egbert and Eisel (Eisel and Egbert 2001). The open source code package provides two 100 $\Omega \cdot m$ uniform half-space time-series data (test1.asc and test2.asc). Each time-series data has five columns of data with a length of 40,000 s. The five columns of data represent the x, y, z direction of the magnetic field and the x, y direction of the electric field for the observed data. The correlation between the signals of the two time-series data provided by the open source package is close to 1 (Egbert and livelybrooks 1996). Furthermore, typical strong interference is added to the original signal of EMTF data (test1.asc) to assess the performance of the OMP method and the proposed method on the time-series. Compared with the apparent resistivity curve by obtained from the remote reference method and the OMP-based overall method.

Figure 5(a) and (b) show the addition of square wave interference on the original time-series of Hx, and the addition of the triangle wave interference on the original time-series of Ey of the EMTF

data for signal-noise separation. At the same time, Figure 5(c) shows the comparison of the apparent resistivity curves obtained by the noisy data, the remote reference method, OMP-based overall method and the proposed method.

(a)

(b)

(c)

Fig. 5 The EMTF data is disturbed by a noisy signal with (a) square wave interference of Hx and (b) charge and discharge triangle wave interference of Ey; (c) is a comparison of the apparent resistivity-phase curves, left figure is ρ_{xy} direction and right figure is ρ_{yx} direction. Among them, curve 1 is original data; curve 2 is noisy data; curve 3 is remote reference method; curve 4 is OMP-based overall method; curve 5 is the proposed method.

Considering that Ey and Hx are correlation channel data, we add the strong interference to the corresponding data. It can be known from Figure 5(a) and 5(b) that the interferences and signals are accurately identified, and the proposed method can extract the profile of the interference. By comparing the original data (Hx and Ey) with the reconstructed data (Hx and Ey) of the proposed method, it is found that the NCC is 0.9583 and 0.9618, respectively, and the frequency spectrum of original data and reconstructed data are also very close. While the NCC is respectively 0.1172 and 0.4316 on the reconstructed data by OMP method, its data basically lose low-frequency useful signal. From Figure 5(c), the apparent resistivity curve 1 of the original data is smooth. Curve 2 is obtained by adding noise to the original data (Hx and Ey), because of the noise added to the relevant channel data, it only causes that the apparent resistivity curve of ρ_{yx} direction appears a variety of change. While curve 3 is obtained from the remote reference method, and uses the EMTF time-series data (test2.asc) as reference data to suppress noise-added data. Since the noise-added data belong to correlated noise, this method makes a few frequency jumps in the low frequency of ρ_{yx} direction, which cannot achieve an ideal effect. Curve 4 is obtained by the OMP-based overall method that can suppress the strong interference in the time-series waveform. The method filters the entire data blindly, causing to the loss of the useful data, the entire apparent resistivity curves still disorder. According to the time-series waveform signal-noise identification and separation, the NCC analysis and the apparent resistivity curves, the shape and smoothness of the apparent resistivity curves obtained by the proposed method that is the curve 5 is close to those of the original data. Thus, the proposed method is more suitable for subsequent measured MT noise separation.

Noise Separation of the Measured MT Data

To verify the effectiveness of the proposed method, Figure 6 shows the signal-noise identification and targeted de-noising effect by comparing with the OMP and the proposed method for the MT signal with square wave and triangle wave interference.

Fig. 6 The signal-noise identification and targeted de-noising for the measured magnetotelluric (MT) data, and (a) is comparison of square wave interference and (b) is triangle wave interference using the OMP method and the proposed method.

In the case of equivalent atoms and the number of iterations, the OMP-based overall method can remove the large-scale strong interference, and some interference still remains, and the useful MT signal segment is also filtered accordingly, resulting in that the reconstructed signal cannot approach the real MT signal. The proposed method carries out signal-noise identification and noise separation, which improves the authenticity of the reconstructed signal, thereby avoiding the under-processing and over-processing of the OMP method.

Results

Apparent Resistivity-Phase Curve Analysis of the Measured Sites

In this section, we compared with the apparent resistivity-phase curves of the original data, the remote reference (RR) method, the orthogonal matching pursuit (OMP) method, the fractal-entropy and clustering method (Li et al. 2018) and the proposed method. These measured sites (D37890 and EL22174) were affected from the square wave and triangle wave interference types in time-series sequence.

Figure 7 shows the comparison of the apparent resistivity-phase curves of the measured MT site (D37890).

Fig. 7 Comparison of apparent resistivity-phase curves of the measured MT site D37890, among them, curve 1 is original data, and curve 2 is the data filtered by the orthogonal matching pursuit (OMP) method, and curve 3 is the result derived from the fractal-entropy and clustering method (Li et al. 2018), and curve 4 is the result derived from the proposed method.

In Figure 7, the apparent resistivity-phase curve of the original data (curve 1) gradually rises to $10^4 \Omega \cdot m$ at 10-0.03 Hz, and the phase curve of the corresponding frequency band is attached to 0° . The curve presents a downward trend at 0.03 Hz, and its corresponding phase jump becomes large. Therefore, the data quality is obviously poor, and shows the near-source effect. By observing the time-series sequence of original data, it is found that strong interference almost submerges the entire low-frequency useful MT signal. As a result, it cannot characterize the underground electrical structure information.

According to the data filtered by the orthogonal matching pursuit (OMP) method, we can see from the curve 2 that the entire low-frequency band has been declining, and the corresponding phase curve is seriously disturbed. Although the de-noising result of time-series waveforms show that large-scale strong interference is also eliminated, it lacks the retention of useful signals and incomplete removal of complex interference, resulting in the failure to provide valid data. The fractal-entropy and clustering method (curve 3) extracts four types of the feature parameters, and has a relatively stable discrimination for the MT signal-noise, and then uses traditional wavelet for targeted de-noising. The curve in the ρ_{yx} direction of 10-0.03 Hz has not been alleviated, which shows a convex state. The difference of characteristic parameters has led to the inability of identify MT signal and noise in this frequency band. Moreover, this method still takes longer time to calculate the feature parameters, and need to predefine different wavelet bases and decomposition layers for wavelet de-noising, and the fractal-entropy and clustering method (Li et al. 2018) cannot obtain the high-quality of MT data. Therefore, the proposed method (curve 4) can purposefully suppress the identified as the interference segments, and preserve the useful MT signals segments, and reconstructed MT signals is close to the natural field MT data. The apparent resistivity-phase curve becomes stable and continuous, and the phase curve has also been greatly improved. The result indicated that the denoised data by the proposed method can more truly reflect the underground electrical structure information.

Figure 8 shows the comparison of the apparent resistivity-phase curves of the measured AMT site (EL22174).

Fig. 8 Comparison of apparent resistivity-phase curves of the measured AMT site EL22174, among them, curve 1 is original data, and curve 2 is the result obtained by the remote reference (RR) method, and curve 3 is the data filtered by the orthogonal matching pursuit (OMP) method, and curve 4 is the result derived from the fractal-entropy and clustering method (Li et al. 2018), and curve 5 is the result derived from the proposed method.

In Figure 8, the apparent resistivity curve rises in the ρ_{xy} direction and drops in the ρ_{yx} direction, and the apparent resistivity curve becomes confusion under the 3 Hz frequency band. The corresponding phase curves are based on 0° and $\pm 180^\circ$. The results show that the data quality of the low frequency band is seriously reduced.

Compared to the curve 2 for the result obtained by the remote reference method, the apparent resistivity-phase curves become smoother, there is still frequency fluctuation at 1-0.3 Hz. Since the remote reference method depends on the selection of reference site and the measured site distance, it is difficult to achieve the desired effect. Curve 3 is the data filtered by the orthogonal matching pursuit (OMP) method that caused to the low-frequency band of the apparent resistivity curve fall seriously, and the corresponding phase curve is more chaotic. Meanwhile, the result cannot provide the reasonable AMT data. Curve 4 is the result derived from the fractal-entropy and clustering method (Li et al. 2018) which obtained the smoother of the entire apparent resistivity curve. Because four characteristic parameters describe the AMT signal and noise at different angles, and the identification of weak noise as a useful signal and a large number of human factors, resulted in the low of AMT data quality in the low frequency band. The proposed method (curve 5) can remove the segment identified as AMT interference, and the denoised data combine with segment identified as AMT signals for the precisely reconstructed data. The apparent resistivity-phase curve is more continuous and smoother. Thus, the result can provide the more reliable and reasonable AMT data for subsequent electromagnetic inversion.

Polarization Direction Analysis

According to the definition of polarization direction of electromagnetic field (Weckmann et al. 2005), it is also an indicator to evaluate the quality of MT/AMT data, and introduced to verify the effectiveness of the proposed method. Figure 9 shows the comparison of electromagnetic polarization direction at 0.3Hz and 4Hz for the measured site EL22174 in the electric and magnetic fields, respectively.

Fig. 9 Comparison of the polarization direction for site EL22174, (a) electric field data at 0.3 Hz, and (b) magnetic field data at 4 Hz.

From Figure 9, the polarization direction of the electric field is derived from the original data, which is concentrated on -80° and 50° at 0.3 Hz. The polarization direction of the original data in the magnetic field is concentrated on -60° and 80° at 4 Hz. Because the time-series sequence of original data contained persistent strong interference. By analyzing the polarization direction of the proposed method, the randomness of the polarization direction of the electromagnetic field increases, and the polarization angle is disordered, which is consistent with the polarization characteristics of the natural field. And combined with the apparent resistivity-phase curve, it has been confirmed that the proposed method can purposefully suppress the strong interference and the reconstructed data is close to natural MT field data.

Discussions

Since the theory of MT sounding was proposed, the problem of noise has been troubling the majority of MT researchers. Typical noise interference is often filled with time domain waveforms, which affects the degradation of data quality in the low frequency band. The apparent resistivity is chaotic and discontinuous, and the polarization direction is highly concentrated in a certain direction. Therefore, the proposed method purposefully eliminates MT strong interference, and improve the quality of low-frequency data in the ending results.

According to the application of existing technology in the MT noise suppression, the reliable MT sounding data is obtained by editing, filtering and signal-noise identification in the time-frequency domain. Therefore, these techniques have provided significantly confusion for the interpretation of geological exploration, and the researches on signal-noise identification has been continuously derived. By extracting the characteristic parameters of MT data for clustering and classification, identifying signals and noise, while the information of low-frequency slow-change of MT is kept abundantly and the data quality is improved. At present, the signal-noise identification technology has extracted a large number of complexity feature parameters and analyzed the essential characteristics of MT data from different angles, such as recurrence analysis, sample entropy, fuzzy entropy, approximate entropy, LZ complexity, fractal dimension, etc. The above-mentioned characteristic parameters are not analyzed for their scale features, only through the fusion of a large number of characteristic parameters, when judging whether a segment of MT signal is affected by noise, the multiple characteristic parameters differ greatly, resulting in the reduction of the accuracy and efficiency of clustering and classification. At the same time, when the MT signal-noise is gradually blurred, the ability of feature recognition is also failed, and the apparent resistivity curve shows a downward trend in the low frequency band, which cannot accurately interpret the deep underground of the electrical structure information.

In this study, when the time scale is uncertain, MSE will extends the sample entropy to multiple time scales, so as to provide additional viewing angles. MSE involves coarse-graining time-series process, which is mainly used to analyze time series at increasingly coarse time resolution. MDE consists of coarse-grain the original signal at different scales, and then obtaining the DE value for the coarse-grained sequences at different scales to obtain the Shannon's entropy characteristics. RCMDE is improved MDE algorithm, using refined composite technology, resulting the better consistency and stability of entropy value, and faster calculation speed. In the MP algorithm, if the vertical projection of the signal (that is residual signal) on the selected atoms is non-orthogonal, it will trim the result of each iteration, and requires many iterations to converge. The OMP algorithm means that all the selected atoms are orthogonality in each step of the decomposition, which makes the OMP algorithm converge faster when the accuracy requirements are the same. Therefore, we mined the robust characteristic parameter RCMDE as the feature parameter for cluster analysis, and divided the MT signal and noise through its scale characteristics with high precision, and used OMP algorithm for noise separation. Through discussing the experimental results, we compare the proposed feature parameters with MSE and MDE, and compare the de-noising method with MP. The proposed method improves MDE and the stability of the distinction between MT signal and noise, and we can use the RCMDE characteristic with the scale factor is 2 to produce two feature parameter values. Using simple clustering and rapidly de-noising method can improve the effect and efficiency of the existing technology in processing MT noisy data. By comparing the remote reference method, OMP-based overall de-noising method and signal-noise identification and separation method (Li et al. 2018). The proposed method effectively improves the multiple frequency information in the low frequency band, and the entire low-frequency curve is smoother and more stable.

In short, many MT signal processing methods eliminate the strong interference of the MT

signal from the perspective of time-frequency domain. The common purpose is to obtain high-quality MT data and provide reliable data for subsequent electromagnetic inversion. In addition, how to refine the multiscale analysis of characteristic parameters and how to optimize the atomic matching in the de-noising method in this paper is still to be studied. Besides, this paper is further research whether the reducing the preset parameters values or optimizing parameters selection can improve MT data processing efficiency and accuracy, and provide richer technical support for the massive MT data. It will inevitably improve the universality of the methods.

Conclusions

We have proposed a novel noise separation method for MT data by using refined composite multiscale dispersion entropy and orthogonal matching pursuit. As a robust feature parameter, RCMDE can generate multiple feature parameter values that distinguish MT signal and noise. OMP is used as a fast de-noising method. In the combination of the RCMDE and OMP, the efficiency and accuracy of MT feature extraction, identification and noise separation are improved. The experimental results show that the identified strong interferences are purposefully eliminated, and the useful MT signals are bounteously reserved, and the quality of MT data is improved. The more continuous and smooth apparent resistivity-phase curve, and the more scattered and random polarization direction obtained by using the proposed method. It will further provide an innovative technology route for MT data processing, and obtain high-precision MT data for subsequent electromagnetic inversion.

Abbreviations

MT: Magnetotelluric; RCMDE: Refined composite multiscale dispersion entropy; MP: matching pursuit; OMP: Orthogonal MP; VMD: Variational mode decomposition; SNR: Signal-noise ratio; MSE: Mean square error; NCC: Normalized cross-correlation; T: Time factor; SE: Sample entropy; FE: Fuzzy entropy; AE: Approximate entropy; DE: Dispersion entropy; RR: Remote reference; FCM: Fuzzy c-mean; MSE: Multiscale entropy; MDE: Multiscale dispersion entropy; DFA: detrended fluctuation analysis; SD: Standard deviation; NCDF: Normal cumulative distribution function.

Authors' contributions

Zhang X wrote the manuscript, designed and analyzed the experiments; Li J and Li DQ conceived the idea and helped revise the manuscript; Li Y helped analyze the experimental results; Liu B and Hu YF helped discuss the algorithm and experiment. All authors have read, revised, and approved the final manuscript.

Funding

This research was supported by the National Key R&D Program of China (No.2018YFC0807802, No.2018YFE0208300), the National Natural Science Foundation of China (No.41874081, No.41404111), the Open Research Fund Program of Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring (Central South University), Ministry of Education (No.2020YSJS06), the Key Laboratory of Geophysical Electromagnetic Probing Technologies of Ministry of Natural Resources (No.KLGEPT201905), and the Natural Science Foundation of Hunan Province (No.2018JJ2258).

Availability of data and materials

The datasets and MATLAB code used during the current study are available from the corresponding authors on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Conflicts of interest

The authors declare no conflict of interest.

Author details

¹ Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment, Monitoring Ministry of Education, School of Geoscience and Info-Physics, Central South University, Changsha 410083, China. ² Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China. ³ Key Laboratory of Geophysical Electromagnetic Probing Technologies of Ministry of Natural Resources, Institute of Geophysical and Geochemical Exploration, Chinese Academy of Geological Science, Langfang 065000, China. ⁴ Hunan Key Laboratory of Nonferrous Resources and Geological Hazard Exploration, Changsha 410083, China. ⁵ College of Mathematics and Physics, Hunan University Arts and Science, Changde 415000, China.

References

- Azami H, Rostaghi M, Abásolo D, Escudero J (2017) Refined composite multiscale dispersion entropy and its application to biomedical signals. *IEEE T Bio-Med Eng* 64: 2872–2879.
- Bandt C, Pompe B (2002) Permutation entropy: a natural complexity measure for time series. *Phys Rev Lett* 88: 174102.
- Becher WD, Sharpe CB (1969) A synthesis approach to magnetotelluric exploration. *Radio Sci* 4(11): 1089–1094.
- Cagniard L (1953) Basic theory of the magnetotelluric method of geophysical prospecting. *Geophysics* 18(3): 605–635.
- Cai JH, Tang JT, Hua XR, Gong YR (2009) An analysis method for magnetotelluric data based on the Hilbert-Huang transform. *Explor Geophys* 40(2): 197–205.
- Costa M, Goldberger AL, Peng CK (2005) Multiscale entropy analysis of biological signals. *Phys Rev E* 71: 021906.
- Egbert GD, Booker JR (1986) Robust estimation of geomagnetic transfer functions. *Geophys J Roy Astr Soc* 87(1): 173–194.
- Egbert GD, Livelybrooks DW (1996) Single station magnetotelluric impedance estimation: coherence weighting and the regression M-estimate. *Geophysics* 61(4): 964–970.
- Eisel M, Egbert GD (2001) On the stability of magnetotelluric transfer function estimates and the reliability of their variances. *Geophys J Int* 144: 65–82.
- Gamble TM, Goubau WM, Clarke J (1978) Magnetotelluric data analysis: removal of bias. *Geophysics* 43(6): 1157–1169.
- Gamble TM, Goubau WM, Clarke J (1979) Magnetotelluric with a remote magnetic reference. *Geophysics* 44(1): 53–68.
- Hermance JF (1973) Processing of magnetotelluric data. *Phys Earth Planet Interiors* 7(3): 349–364.
- Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc R Soc A Math Phys Eng Sci* 454: 903–995.
- Jin W, Wang L, Zeng X, Liu Z, Fu R (2014) Classification of clouds in satellite imagery using over-complete dictionary via sparse representation. *Pattern Recogn Lett* 49(1): 193–200.
- Jones AG, Chave AD, Egbert GD, Auld D, Bahr K (1989) A comparison of techniques for magnetotelluric impedance estimation. *J Geophys Res* 94(10): 14201–14213.
- Li G, Liu XQ, Tang JT, Deng JZ, Hu SG, Zhou C, Chen CJ, Tang WW (2020) Improved shift-invariant sparse coding for noise attenuation of magnetotelluric data. *Earth Planets Space* 72: 45.

- Li J, Zhang X, Gong JZ, Tang JT, Ren ZY, Li G, Deng YL, Cai J (2018) Signal-noise identification of magnetotelluric signals using fractal-entropy and clustering algorithm for targeted de-noising. *Fractals* 26(2): 1840011.
- Li J, Zhang X, Tang JT (2020) Noise suppression for magnetotelluric using variational mode decomposition and detrended fluctuation analysis. *J Appl Geophys* 180: 104127.
- Mallat SG, Zhang Z (1993) Matching pursuit with time-frequency dictionaries. *IEEE Trans Signal Process* 41(12): 3397–3415.
- Mitiche I, Morison G, Nesbitt A, Hughes-Narborough M, Stewart BG, Boreham P (2018) Classification of partial discharge signals by combining adaptive local iterative filtering and entropy features. *Sensors* 18: 406.
- Pati YC, Rezaiifa R, Krishnaprasad PS (1993) Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. *Proceedings of 27th Asilomar Conference on Signals Systems and Computers* 40–44.
- Qi J, Zhang L, Zhang K, Li L, Sun J (2020) The application of improved differential evolution algorithm in electromagnetic fracture monitoring. *Adv Geo-energ Res* 4: 233–246.
- Rostaghi M, Azami H (2016) Dispersion entropy: A measure for time series analysis. *IEEE Signal Proc Let* 23: 610–614.
- Tang JT, Li J, Xiao X, Zhang LC, Lv QT (2012) Mathematical morphology filtering and noise suppression of magnetotelluric sounding data. *Chin J Geophys* 55(5): 1784–1793.
- Tikhonov AN (1950) On determining electrical characteristics of the deep layers of the Earth's crust. *Dokl. Akad. Nauk SSSR* 73: 295–297.
- Trad DO, Travassos JM (2000) Wavelet filtering of magnetotelluric data. *Geophysics* 65(2): 482–491.
- Tropp JA, Gilbert AC (2007) Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans Inf Theory* 53(12): 4655–4666.
- Vallianatos F (1996) Magnetotelluric response of a randomly layered earth. *Geophys J Int* 125(2): 577–583.
- Wang H, Campaña J, Cheng JL, Zhu GW, Wei WB, Jin S, Ye GF (2017) Synthesis of natural electric and magnet Time-series using Inter-station transfer functions and time-series from a Neighboring site (STIN): Applications for processing MT data. *J Geophys Res-Sol Ea* 122(8): 5835–5851.
- Wang JB, Wang SX, Yin HJ, Zhang R (2013) A self-adaption denoising method using orthogonal matching pursuit. *SEG Technical Program Expanded Abstracts*.
- Weckmann U, Magunia A, Ritter O (2005) Effective noise separation for magnetotelluric single site data processing using a frequency domain selection scheme. *Geophys J Int* 161(3): 635–652.
- Xu Y, Chen R, Li Y, Zhang P, Yang J, Zhao X, Liu M, Wu D (2019) Multispectral image segmentation based on a fuzzy clustering algorithm combined with tsallis entropy and a gaussian mixture model. *Remote Sens* 11: 2772.
- Zhang H, Liu J, Chen L, Chen N, Yang X (2019) Fuzzy Clustering algorithm with non-neighborhood spatial information for surface roughness measurement based on the reflected aliasing images. *Sensors* 19: 3285.
- Zhang YD, Tong SG, Cong FY, Xu J (2018) Research of feature extraction method based on sparse reconstruction and multiscale dispersion entropy. *Appl Sci* 8: 888.