Journal of Neuroscience Methods 295 (2018) 111–120

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Cosine similarity as a sample size-free measure to quantify phase clustering within a single neurophysiological signal

Elizabeth P. Chou[a], Shen-Mou Hsu[b,*]

[a] Department of Statistics, National Chengchi University, Taipei, Taiwan, ROC
[b] Imaging Center for integrated Body, Mind, and Culture Research, National Taiwan University, Taipei, Taiwan, ROC

HIGHLIGHTS

- The existing measures of phase clustering suffer from sample size bias.
- Cosine similarity (CS) is robust against sample size variation.
- CS could detect inherent nature of phase clustering between datasets.

ARTICLE INFO

Article history:
Received 29 September 2017
Received in revised form 5 December 2017
Accepted 11 December 2017
Available online 13 December 2017

Keywords:
Cosine similarity
ITC
Oscillations
Phase
Phase clustering

ABSTRACT

Background: Phase clustering within a single neurophysiological signal plays a significant role in a wide array of cognitive functions. Inter-trial phase coherence (ITC) is commonly used to assess to what extent phases are clustered in a similar direction over samples. However, this measure is especially dependent on sample size. Although ITC was transformed into ITCz, namely, Rayleigh's Z, to "correct" for the sample-size effect in previous research, the validity of this strategy has not been formally tested.

New method This study introduced cosine similarity (CS) as an alternative solution, as this measure is an unbiased and consistent estimator for finite sample size and is considered less sensitive to the sample-size effect.

Results: In a series of studies using either artificial or real datasets, CS was robust against sample size variation even with small sample sizes. Moreover, several different aspects of examinations confirmed that CS could successfully detect phase-clustering differences between datasets with different sample sizes.

Comparison with existing methods Existing measures suffer from sample-size effects. ITCz produced a mixed pattern of bias in assessing phase clustering according to sample size, whereas ITC overestimated the degree of phase clustering with small sample sizes.

Conclusions: The current study not only reveals the incompetence of the previous "correction" measure, ITCz, but also provides converging evidence showing that CS may serve as an optimal measure to quantify phase clustering.

1. Introduction

A growing body of evidence shows that phases of brain oscillations, as revealed by neurophysiological signals, play a significant role in neural coding and communication (Sauseng and Klimesch, 2008; van Rullen et al., 2011). In principle, there are two qualitatively different approaches to study phase information. One focuses on examining the phase relationship between *two* signals from separate sensors or brain areas, such as the index "phase-locking value" (Lachaux et al., 1999), whereas the other focuses on examining how phases behave within a *single* signal. A standard tool in the *latter* approach relies on assessing to what extent phases are clustered in a similar direction over samples (e.g., trials or time points) to determine whether a "preferred" angle is present in the signal. To date, phase clustering is associated with a wide array of cognitive functions, including attention (Lakatos et al., 2008), conscious perception (Palva et al., 2005), music perception (Doelling and Poeppel, 2015), object recognition (Tallon-Baudry et al., 1996), speech discrimination (Luo and Poeppel, 2007), reaction time (Drewes and van Rullen, 2011), and working memory (Bonnefond and Jensen,

* Corresponding author at: Imaging Center for Integrated Body, Mind, and Culture Research, National Taiwan University, 1F No.49, Fanglan Rd Da'an Dist, Taipei City 106, Taiwan ROC.
E-mail address: smhsu@ntu.edu.tw (S.-M. Hsu).

https://doi.org/10.1016/j.jneumeth.2017.12.007
0165-0270/

Fig. 1. A schematic diagram illustrating the computation of (A) ITC and (B) CS over 3 phase samples. ITC computes the mean resultant length of phase vectors on a unit complex plane, whereas the CS measure computes the mean cosine angle of all given pairs of phase vectors.

2012). Phase clustering is even thought to serve as a potential source for event-related potentials (Makeig et al., 2002).

In most previous research, inter-trial phase coherence (ITC or phase-locking factor (Tallon-Baudry et al., 1996)) was commonly used to quantify phase clustering. However, the ITC measure is problematic because it especially depends on sample size (Edwards et al., 2009). This major problem arises because ITC computation first involves time-frequency analysis of neurophysiological signals to derive a phase vector for each sample on a unit complex plane. Next, ITC values are computed by averaging the lengths of phase vectors across samples (Fig. 1A). However, ITC is a biased and consistent estimator. An estimator is unbiased if its expected value $E\ \hat{I}\ $ equals I, but ITC is estimated as $\hat{I} = E(e^{i\theta_j})$ (see the "Materials and methods" section) and is thereby biased because $E\ \hat{I}\ \geq I$, and it is consistent when $\lim_{n\to\infty} \hat{I} = I$ and $\lim_{n\to\infty} Var(\hat{I}) = 0$ (see (Kutil, 2012) for the mathematical proof). To see this bias, consider phase vectors randomly scattered around a complex plane with an expected value of zero. During ITC computation, it is unlikely that these vectors could be perfectly arranged to sum to 0 due to variance. As a result, the ITC measure suffers from a sample-size bias.

In practice, ITC values are calculated for each experimental condition in which trial samples are grouped according to associated behavioral responses or task demands. The sample-size bias present in the ITC measure may thus have a deleterious impact on interpretation of the results, particularly when ITC is compared between conditions that greatly differ in trial count. Although the optimal situation is to design an experiment in which trial numbers are balanced between conditions, this is not always possible. For example, during the preprocessing procedure in magnetoencephalography (MEG) analyses, the number of trials surviving artifact rejection may vary across experimental conditions. In addition, the behavioral outcomes may not be under the control of the experimenter. For example, in a typical conscious perception experiment, even though a stimulus is presented at threshold, the proportions of consciously detected and undetected trials may still vary as a result of fluctuations in participants' performance from moment to moment.

Several strategies have been employed to address this issue. One straightforward strategy is to randomly delete trials from an experimental condition with a greater number of trials such that the trial numbers are matched between the two conditions being compared. Then, ITC values can be assessed between conditions with equal trial counts. This procedure could be performed in combination with a bootstrap method to ensure reliability (Hsu and Yang, 2017 in press; van Diepen et al., 2015). Here, we refer to this trial reduction in combination with the bootstrapping approach as ITCe. Unfortunately, the ITCe measure is computationally demanding and consequently time-consuming. In addition, ITCe comes at the cost of decreasing statistical power due to fewer trials being used. Moreover, in some situations, such as developmental studies with infants wherein data are scarce, this strategy becomes particularly problematic and impractical.

A different line of reasoning is to keep original trials while performing some "correction" methods to mitigate the sample-size effect. One widely adopted approach is to transform ITC to ITCz, namely, Rayleigh's Z (Fisher, 1993). This procedure involves weighting original ITC values with trial number to correct the bias. Despite its current application (Bonnefond and Jensen, 2012; Cohen, 2014; Samaha et al., 2015), the validity of this strategy has surprisingly not been formally tested.

In addition to the ITCz measure, cosine similarity (CS) may provide another potential solution to "correct" for the sample-size effect. CS is commonly used to measure cohesion within vectors in the fields of text mining (Baeza-Yates and Ribeiro-Neto, 1999; Hotho et al., 2005), machine learning (Pang-Ning et al., 2006) and even rhythmic neuronal synchronization (Vinck et al., 2010). Specifically, this measure computes the similarity of any two vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ in terms of their cosine angle θ using the following equation:

$$cos\theta_{\mathbf{v}_i\mathbf{v}_j} = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{||\mathbf{v}_i||||\mathbf{v}_j||}$$

The zero value indicates that the vectors are orthogonal and have low similarity, −1 indicates that the vectors are totally opposite, and 1 indicates that the vectors are highly similar and point in the exact same direction. Given that phases of neurophysiological signals can be represented in terms of vectors on a complex plane, we reason that the CS measure allows quantification of phase clustering by computing the mean cosine angle of all given pairs of phase vectors (Fig. 1B). In other words, phase clustering is assessed by examining how similar the phase vector observed in one data sample is to the phase vectors observed in other data samples. As a result, a CS value close to 1 would reflect that phase vectors across samples are highly similar and thus converge around a similar direction (i.e., high phase clustering). More importantly, unlike ITC, the expected value of the estimated CS is equal to CS, i.e., $E\ \hat{CS}\ = CS$, and $\lim_{n\to\infty} Var\ \hat{CS}\ = 0$. To demonstrate that $\hat{CS}$ is an unbiased estimator of CS, we can obtain

$$E\ \hat{CS}\ = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E(\cos\ \theta_i - \theta_j\ ) = CS$$

Therefore, $\hat{CS}$is an unbiased estimator of *CS*. To demonstrate that $\hat{CS}$ is a consistent estimator of *CS*, we can obtain

$$Var(\hat{CS}) = Var(\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \cos(\theta_i - \theta_j)) = \frac{4}{n^2(n-1)^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Var(\cos(\theta_i - \theta_j)) = \frac{4}{n^2(n-1)^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E(\cos(\theta_i - \theta_j)^2) - E(\cos(\theta_i - \theta_j))^2$$

$$\lim_{n\to\infty} P(|\hat{CS} - CS| < \varepsilon) = 1, \quad \forall \varepsilon > 0$$

Therefore, CS is a consistent estimator because $\lim_{n\to\infty} Var\ \hat{CS} =$ 0 and E $\hat{CS} \to CS$ *as* $n \to \infty$. Therefore, mathematically speaking, the CS measure is an unbiased and consistent estimator for finite sample size and is considered less sensitive to the sample-size effect.

In this article, CS is introduced as an alternative method to compute phase clustering and resolve the sample-size effect. We first systematically examined the validity of the existing correction measure ITCz as well the traditional measure ITC. Then, we evaluated the performance of the CS measure against that of other measures to probe the optimal measurement of phase clustering. In the context of both artificially constructed and real MEG datasets, we demonstrate that compared with previous measures, CS may serve as a better phase-clustering measure to "correct" for sample-size effects and indicate its practicality for analyzing neurophysiological data.

2. Materials and methods

2.1. Datasets

2.1.1. Study I: phase-clustering measures as a function of sample size

Artificial data were created using the CircStat toolbox (Berens, 2009) by randomly sampling phase samples (- to) from a von Mises distribution, which is the circular analog of a normal distribution. Both sample size (size 1–100) and levels of (= 0.01, 0.1, 0.2, 0.5) were manipulated. The parameter indicates the dispersion of the von Mises distribution. As approaches 0, the data points become uniformly distributed on the phase circle. After the sampling procedure, the phase-clustering values of each measure (i.e., ITC, ITCz, and CS) were computed across samples, yielding a total of 100 (sample size) ×4 ( ) datasets. To ensure reliability, the above procedure was repeated 1000 times. For each of the 1000 iterations, the von Mises distribution had a randomly selected mean direction. Finally, the phase-clustering values were averaged over iterations and subjected to statistical analysis.

Study I-2 used a previous experimental dataset to examine how the phase-clustering measures performed in actual data (Hsu and Yang, 2017 in press), particularly given that in real situations, phases comprise different frequency ranges. In our previous experiment, each trial began with the presentation of a fixation cross for 800–1000 ms, followed by a forward mask for 300 ms, a stimulus for 17 ms, and a backward mask for 33 ms. After a blank for 250 ms, a response window with three options was displayed. Participants had up to 3000 ms to report (1) that they could recognize the facial identify of the target stimulus by selecting the option "Liu" (face identification hit trials, FI), (2) that they could detect the presence of the target face but could not recognize the facial identity of the target by selecting the option "Face" (face detection hit trials, FD), or (3) that they could not see a face by selecting the option "No" (face miss trials, FM). Thirteen participants participated in the experiment (8 males, mean age ± STD = 26.38 ± 3.23 years, range = 21–31), while MEG signals were recorded (Yokogawa, Co., Tokyo). The signals were digitized at 1000 Hz and filtered with 0.3-Hz high-pass and 500-Hz low-pass cutoffs and a 60-Hz notch. Continuous MEG data were segmented into 2000-ms epochs starting from 1000 ms before the onset of the forward mask. Trials contaminated with eye movements, eye blinks, and muscular artifacts were rejected via visual inspection. Time-frequency representations of phase information in the MEG signals were computed using Morlet's wavelets (m = 7) on every sensor, frequency (8–100 Hz, step: 2 Hz) and time point (step: 10 ms) in each trial. Fieldtrip (Oostenveld et al., 2011) and MATLAB (MathWorks, Natick, MA) software was used for data processing.

Because our previous results indicated high ITC values in FI trials, for every participant, we selected the peak data points at 400 ms after first mask onset, the sensor AG098 and all frequency ranges in FI trials for the analyses. Next, real datasets were created by randomly drawing samples from the above data pool with sample sizes ranging from 1 to 30 (the minimum number of FI trials among participants). The phase-clustering values of each measure were then computed across samples, leading to 30 (sample size) ×47 (frequency) datasets for each participant. We repeated the sampling procedure 1000 times and averaged the phase-clustering values across iterations and participants for further analysis.

2.1.2. Study II: validity of phase-clustering measures in simulation studies

To validate whether each measure could successfully differentiate datasets with *different* degrees of phase clustering when sample sizes differ between two datasets, we deliberately constructed two artificial datasets in Study II-1. In one dataset (data20_low), 20 data points were randomly sampled from a von Mises distribution with = 0.01. In the other dataset (data100_high), 100 data points were randomly sampled from a von Mises distribution with = 0.2. As such, phases were de facto better aligned in "data100_high" due to the higher concentration of the von Mises distribution (see Section 3.1 for details).

To further validate whether each measure could reflect that in reality, the two datasets had *the same* degree of phase clustering over samples when sample sizes differed between datasets, we additionally constructed two datasets drawn from the *same* von Mises distribution with = 0.2 in Study II-2. One dataset had a sample size of 20 (data20_same), and the other had a sample size of 100 (data100_same). As a result, these two datasets had the same phase clustering but differed in sample sizes.

For Study II-1 and Study II-2, the sampling procedure was repeated 1000 times to generate sampling distributions for statistical testing. For each of the 1000 iterations, each type of phase-clustering values was computed based on the samples drawn from the von Mises distribution with a randomly selected mean direction.

2.1.3. Study III: validity of phase-clustering measures in a real experimental setting

The finding from our previous experiment, as described in Study I, was used as a benchmark to evaluate how the phase-clustering measures performed in a real experimental setting. In this previous finding, stronger phase clustering was found in FI trials than in FM trials. This significant effect occurred at 10 Hz in the right frontal-parietal-temporal sites between 240 and 580 ms after the onset of

the first mask. Because there were fewer FI trials (mean number of trial ± SD = 55 ± 6) than FM trials (96 ± 14) for 12 of 13 participants, in a previous analysis, we employed ITCe (i.e., the trial-reduction analysis approach in combination with the bootstrap procedure) to address the issue of unbalanced trial counts. Specifically, the trials were randomly selected without replacement from the condition with a greater number of trials such that the trial numbers were matched between the two conditions being compared. This procedure was repeated 2000 times, and for each iteration, an ITCe value was computed. As a result, a distribution of ITCe was generated for every sensor-time-frequency point. The obtained distribution was then characterized by its mean, which was submitted to cluster-based permutation tests (see Section 2.3.3).

2.2. Phase-clustering measure

2.2.1. ITC measure

ITC values were computed by the following equation:

$$ITC = |\frac{1}{n}\sum_{j=1}^{n} z_j|$$

Here, n denotes the sample size. z_j is a complex vector representing a given sample in a time-frequency domain after normalization to unity as follows:

$$z_j = e^{i\theta_j} = \cos\theta_j + isin\theta_j$$

where $j = 1, \ldots n$, and θ_j is the phase of the sample z_j. Accordingly, ITC can be construed as totaling the straight-line distance from the starting point to the ending point of samples and dividing by the sample size. In other words, ITC indicates the mean resultant length of samples. The ITC values are bounded between 0 and 1. An ITC close to 0 reflects low phase clustering (i.e., the distribution of phase angles across samples is uniform), whereas an ITC close to 1 reflects strong phase clustering (i.e., all samples exhibit the same phase). For all artificial data, the complex vector z_j was derived by transforming phase samples from von Mises distributions using Euler's formula. For all real data, complex vectors were derived from wavelet analysis of every sensor, frequency and time point in each trial, as previously described.

2.2.2. ITCz measure

To compute ITCz (aka, Rayleigh's Z), we transformed the obtained ITC values using the following formula:

$$ITCz = n \times ITC^2$$

where n is the sample size. Therefore, unlike ITC, ITCz values can be larger than 1.

2.2.3. CS measure

For any two samples ($z_i = e^{i\theta_i} = \cos\theta_i + i\sin\theta_i$ and $z_j = e^{i\theta_j} = \cos\theta_j + i\sin\theta_j$) on a unit complex plane, similarity between these two samples can be represented in terms of their cosine angle:

$$\cos\ \theta_i - \theta_j\ = \frac{e^{i(\theta_i-\theta_j)} + e^{-i(\theta_i-\theta_j)}}{2} = \cos\ _i\cos\ _j + \sin\ _i\sin\ _j$$

To obtain CS values, we computed the cosine angle of all possible sample pairs as follows:

$$CS = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \cos\ \theta_i - \theta_j$$

Here, the constant $\frac{2}{n(n-1)}$ is the number of sample pairs being measured. A CS close to 0 reflects low phase clustering, whereas a CS close to 1 reflects strong phase clustering. CS may also yield negative values when the sample size is small due to high variance.

2.3. Statistical analysis

2.3.1. Study I: phase-clustering measures as a function of sample size

Slope tests were performed for both artificial and real datasets to examine whether the values of each phase-clustering measure changed over sample size as follows:

$$t = \frac{\hat{\beta}_1}{\frac{\sum_{i=2}^{n} I_i - \hat{I}^{\ 2}}{n-2} \Big/ \sum_{i=2}^{n} N_i - \bar{N}^{\ 2}}$$

I denotes the phase-clustering value of a given dataset. N represents the sample size of the dataset. $\hat{\beta}_1$ is the estimated slope:

$$\hat{\beta}_1 = \frac{\sum_{i=2}^{n} I_i N_i - \frac{\sum_{i=2}^{n} I_i \sum_{i=2}^{n} N_i}{n}}{\sum_{i=2}^{n} N_i - \bar{N}^{\ 2}}$$

If a given dataset is susceptible to sample-size effects, the amount of change in I should be statistically significant over sample size N. If not, we would expect the regression line to be horizontal, i.e., a zero slope.

To complement the slope tests (particularly the null effects), Bayes factors were calculated to determine how much evidence exists in favor of the null hypothesis, i.e., phase-clustering values did not change across sample sizes, as follows:

$$B_{10} = \frac{P(D|H_1)}{P(D|H_0)}$$

Here, D in the numerator and denominator respectively denotes the phase-clustering values under the alternative and the null hypotheses. Evidence of different hypotheses can be calculated in terms of marginal likelihoods. The ratio of these likelihoods is a Bayes factor. During the computation implemented by R software (http://www.R-project.org.) with the "BayesFactor" package (Morey and Rouder, 2015), the choice of a prior distribution is subjective. In practice, a Standard Cauchy prior width of r = 1 was used as the scale of prior distribution for computing Bayes factors (Jeffreys, 1961). To draw conclusions, a conventional cut-off was used (Jeffreys, 1961). A Bayes factor close to one indicates that the data are equally consistent with both the null and alternative hypotheses. A value greater than one indicates increasing evidence for the alternative hypothesis, and values approaching zero indicate increasing evidence for the null.

2.3.2. Study II: validity of phase-clustering measures in simulation studies

In Study II-1, independent, two-sample, two-tailed t-tests were conducted to examine whether the sampling distribution of each type of phase-clustering values in "data20_low" differed from that in "data100_high". In Study II-2, Bayes factors, as previously described, were calculated for each phase-clustering measure to determine how much evidence exists in favor of the null hypothesis, i.e., the same degree of phase clustering between "data20_same" and "data100_same".

2.3.3. Study III: validity of phase-clustering measures in a real experimental setting

Similar to our previous study, cluster-based permutation tests were employed for statistical testing while controlling for multiple comparisons to determine whether phase clustering derived from each measure differed significantly between FI and FM trials. In the

tests implemented by Fieldtrip software (Oostenveld et al., 2011), phase-clustering differences between trial groups were quantified by means of paired *t*-tests for every sensor-time-frequency sample. The samples with t values exceeding the threshold ($p < 0.05$) were then clustered in connected sets based on spatial, temporal or frequency adjacency with a minimum of two neighborhood sensors. The cluster with the maximum sum of t values was used as a test statistic. A distribution was generated by randomly permuting the data across the conditions for each participant and then recalculating the test statistic 1000 times using a Monte Carlo estimate. Finally, p values (two-sided, $p < 0.05$) were determined by evaluating the proportion of the distribution resulting in a test statistic larger than the observed statistic.

To quantify and compare the statistical results among the phase clustering measures, we used our previous finding as a benchmark and evaluated how each measure performed against the benchmark finding. Two indices, sensitivity and specificity, which were defined according to signal detection theory, were adopted for this evaluation (Grandchamp and Delorme, 2011). Sensitivity was adopted to assess to what extent each measure could "accurately" detect the *significant* results as indicated by the benchmark finding, whereas specificity assessed to what extent each measure could "accurately" detect the *non-significant* results as indicated by the benchmark finding. These two indices can be formalized as follows:

$$Sensitivity = \frac{\mathrm{TP}}{(TP + FN)}$$

$$Specificity = \frac{\mathrm{TN}}{(FP + TN)}$$

Here, true positive (TP) represents the number of significant sensor-time-frequency points identified by a given phase-clustering measure inside the benchmark finding; false positive (FP) represents the number of significant sensor-time-frequency points outside the benchmark finding; false negative (FN) represents the number of non-significant sensor-time-frequency points inside the benchmark finding; and true negative (TN) represents the number of non-significant sensor-time-frequency points outside the benchmark finding.

To better evaluate the validity of phase clustering measures, receiver operating characteristic (ROC) analysis was additionally performed. In previous cluster-based permutation tests, we specified a minimum number of 2 neighborhood sensors required for a selected sample (i.e., a sample with a t-value exceeding the threshold) to be included in the clustering algorithm. For this analysis, we varied this criterion by changing the parameter from 1 to 4 neighborhood sensors as these 4 levels of parameters produced significant results. Therefore, four hit (aka, *Sensitivity*) and false alarm rates (1-*Specificity*) could be computed and used to construct a ROC curve. The area under the ROC curve (AUC) was then obtained for each of the three measures. The obtained value represents how well a given measure could predict the benchmark finding, with a chance level of 0.5 and a maximal performance of 1.

3. Results

3.1. Study I: phase-clustering measures as a function of sample size

Study I examined how phase-clustering measures, CS, ITC and ITCz, varied as a function of sample size in either a well-controlled context (Fig. 2, left panel) or in a real situation in which phases comprised different frequency ranges (Fig. 2, right panel). For each sample size, κ (the level of distribution dispersion) or frequency range, the phase-clustering values of measure were computed across samples randomly drawn from a von Mises distribution or from our previous experimental data (see Section 2.1.1 for details). This sampling process was repeated 1000 times to ensure reliability. Next, the values were averaged across 1000 iterations in artificial datasets and across 1000 iterations and 13 participants in real datasets for the following analyses.

For both artificial and real datasets, ITC values were high with small sample sizes but gradually decreased when sample size increased. Visual inspection also revealed that the values continued to drop even when sample size reached a reasonable number, such as 30 samples. This pattern of results persisted regardless of the levels of dispersion of the von Mises distributions used in artificial datasets (slope tests on the slope changes, all p values <0.001; Fig. 2A, left panel) or levels of frequency ranges in real datasets (all p values <0.001; Fig. 2A, right panel). Additional analysis provided further support for the latter report, showing that even for each participant, the ITC values also varied according to sample size at all frequency ranges (all p values <0.001). Moreover, notably, as revealed by the absolute values of the slope estimates, the sample-size effects became much more evident with lower levels of distribution dispersion ($=0.01$:*abs* $\hat{\beta}_1$ $=0.0025376>$ $=0.1$: *abs* $\hat{\beta}_1$ $=0.0024405>$ $=0.2$:*abs* $\hat{\beta}_1$ $=0.0021842>$ $=0.5$:*abs* $\hat{\beta}_1$ $=0.0014491$); however, no clear pattern was observed between the sample-size effects and frequency ranges.

ITCz values also significantly differed across sample sizes for all levels of distribution dispersion (all p values <0.001; Fig. 2B, left panel) and frequency ranges (all p values <0.05; Fig. 2B, right panel; for individual participants, over 90% of the frequency points with p values <0.05). However, distinct from ITC, ITCz produced a distinct pattern of results according to sample size for artificial datasets. The ITCz values were drastically enhanced when the sample size increased, and this phenomenon was evident with higher levels of distribution dispersion ($=0.01$: *abs* $\hat{\beta}_1$ $=3.064\text{e}{-04}<$ $=0.1$: *abs* $\hat{\beta}_1$ $=0.0019557<$ $=0.2$:*abs* $\hat{\beta}_1$ $=0.0086609<$ $=0.5$:*abs* $\hat{\beta}_1$ $=0.0569799$). For actual datasets, the results were somewhat mixed. ITCz values decreased with increasing sample size in some frequency ranges (14–24, 34–50, 54, 58–72, 76–78, and 98–100 Hz), whereas the values exhibited the opposite pattern in the remaining frequency ranges.

For the CS measure, a distinct pattern was observed. In general, the CS values did not change according to sample size for both artificial datasets (all p values >0.1, except for $=0.2$, $p=0.005$; Fig. 2B, left panel) and real datasets (all p values >0.05, except for the datasets at 32, 40, 60, 62, 70, 74, 82, 84, 86, and 96 Hz, with p values <0.05). To further examine these null effects, Bayes factor tests were additionally conducted. Consistent with previous analyses, the obtained values indicated strong evidence that the CS results favored the null hypothesis (artificial dataset: $=0.01$,$B_{10}=0.27$; $=0.1$,$B_{10}=0.37$; $=0.2$,$B_{10}=5.74$, $=0.5$,$B_{10}=0.12$; real dataset: all $B_{10}<1$ with a mean = 0.35, except for the datasets at 20, 32, 40, 48, 50, 60, 62, 70, 74, 82, 84, 86, 96 Hz with $B_{10}=1.15$, 1.42, 1.48, 1,12, 1.15, 77.70, 2.50, 3.70, 16.37, 4.06, 1.42, 36.66, 15.46). A closer visual inspection revealed strong fluctuations in CS values with extremely small sample sizes. Given this, we performed the tests on the values with a sample size greater than 11 and 8, respectively, for those artificial and real datasets with $B_{10}>1$. The null hypothesis was now supported for all frequency ranges (all $B_{10}<1$ with a mean = 0.45). We also confirmed that this pattern of results persisted in the real datasets at the level of individual participants (more than 90% of the frequency points with $B_{10}<1$ with a mean = 0.32 after excluding the values with a sample size less than 13).

Fig. 2. (A) ITC, (B) ITCz and (C) CS values as a function of sample size. The left panel shows the phase-clustering values derived from artificial phase samples that were randomly drawn from a von Mises distribution with different sample sizes and levels of dispersion. The right panel shows the phase-clustering values derived from real phase samples that were randomly drawn from our previous experimental data with different sample sizes and frequency ranges. The values represent the mean over 1000 sampling iterations for artificial datasets and the mean over 1000 iterations and 13 participants for real datasets.

3.2. Study II: validity of phase-clustering measures in simulation studies

In practice, whether phase-clustering differences between two given datasets can be accurately detected is the major focus of most research. Therefore, Study II aimed to validate to what extent each measure could reveal the inherent nature of phase clustering between datasets with different sample sizes. For this purpose, the following possible scenarios could be considered: (1) a dataset with *higher* phase clustering but *smaller* sample size relative to a dataset with *lower* phase clustering but *larger* sample size, (2) a dataset with *lower* phase clustering but *smaller* sample size relative to a dataset with *higher* phase clustering but *larger* sample size and (3) two datasets with the *same* degree of phase clustering but *differing* sample sizes. Given that the first scenario is redundant because it does not allow us to properly address the issue of the sample-size bias present in the ITC measure (i.e., high phase clustering for small sample size), we considered only the latter two scenarios in Study II.

In Study II-1, we first examined which measure could better detect the true *difference* in phase clustering between two datasets that had different degrees of phase clustering but different sample sizes (i.e., the second scenario). Two artificial datasets, "data20_low" and "data100_high", were deliberately constructed. The former had 20 samples drawn from a von Mises distribution with = 0.01, whereas the latter had 100 samples drawn from a von Mises distribution with = 0.2. Therefore, "data20_low" with a smaller sample size had lower phase clustering relative to "data100_high" with a larger sample size due to the nature of indicated in Fig. 2. For both datasets, we computed the phase-clustering values of each measure across randomly drawn samples and repeated this procedure 1000 times to generate two sampling distributions for statistical testing (Fig. 3, left panel).

Not surprisingly, in opposition to the known fact, the ITC measure revealed significantly higher phase clustering in "data20_low" (mean value $\pm$ SD = 0.20 ± 0.10) than that in "data100_high" (0.12 ± 0.06; independent two-sample *t*-test on mean difference, $t(1632.7) = 20.673$, $p < 0.001$; Fig. 3A, left panel). This finding reflected a sample-size bias in which phase clustering indexed by ITC in "data20_low" was overestimated due to small sample sizes. By contrast, both the ITCz and CS measures revealed the known fact and showed significantly higher phase clustering in "data100_high" (ITCz: 1.91 ± 1.70, CS: 0.0091 ± 0.02; Fig. 3B and C, left panel) than that in "data20_low" (ITCz: 1.01 ± 0.94, CS: 0.0006 ± 0.05; ITCz: $t(1557.8) = 14.612$, $p < 0.001$, CS: $t(1236.2) = 5.2046$, $p < 0.001$).

A complementary approach to validate phase-clustering measures is to probe to what extent each measure could better detect the *similarity* of phase clustering between datasets that have the same degree of phase clustering but differing sample sizes (i.e., the third scenario). In Study II-2, we examined two deliberately

Fig. 3. The validity of the (A) ITC, (B) ITCz and (C) CS measures in detecting datasets with different (left panel) or the same degree (right panel) of phase clustering. The left panel shows the density distributions of the phase-clustering values of two datasets, "data20_low" and "data100_high". The distribution of "data20_low" was created by computing values across 20 phase samples that were randomly drawn with 1000 repetitions from a von Mises distribution with =0.01. The distribution of "data100_high" was created by computing values across 100 samples that were randomly drawn with 1000 repetitions from a von Mises distribution with =0.2. Therefore, "data20_low" inherently had lower phase clustering than did "data100_high". The right panel shows the density distributions of the phase-clustering values of the datasets "data20_same" and "data100_same". The distributions were created by respectively computing values across 20 or 100 phase samples that were randomly drawn with 1000 repetitions from the same von Mises distribution with =0.2. Therefore, both datasets inherently had the same degree of phase clustering.

constructed artificial datasets "data20_same" and "data100_same". Although the former had 20 samples, and the latter had 100 samples, all samples were randomly drawn from the same von Mises distribution and thereby had an identical degree of phase clustering. As before, the sampling procedure was repeated 1000 times, and for each iteration, the phase-clustering values were computed across the samples, yielding two sampling distributions (Fig. 3, right panel).

As expected, the ITC measure could not detect the true nature of phase clustering in the datasets. Different from the known fact, higher phase clustering was found in "data20_same" (0.22 ± 0.11) than in "data100_same" (0.13 ± 0.06; Fig. 3A, right panel). A Bayes factor also indicated strong evidence that the ITC results supported this alternative hypothesis ($B_{10} = 4.67e+97$) rather than the null hypothesis (i.e., the known fact). For the ITCz measure, higher phase clustering was similarly found in "data100_same" (1.99 ± 1.68) compared to "data20_same" (1.22 ± 1.16), which also provided strong evidence for the alternative hypothesis ($B_{10} = 4.75e+97$). However, distinct from the above two measures, the CS measure accurately revealed comparable degrees of phase clustering between the two datasets and was in favor of the null hypothesis ("data20_same": 0.64 ± 0.023, "data100_same": 0.64 ± 0.005; $B_{10} = 0.04$). Taken together, the overall results showed that when sample sizes differed, the CS measure exhibited better sensitivity to detect the inherent nature of phase clustering between the datasets, regardless of whether these datasets had different or the same degree of phase clustering.

3.3. Study III: validity of phase-clustering measures in a real experimental setting

Although Study I and II addressed the main issue regarding to what extent these phase-clustering measures are immune to the sample-size effect, how each of these measures perform in real situations in which the results are assumed is unclear. In other words, could these measures detect proper or "accurate" data points that have been previously shown significant phase clustering? Using the trial-reduction analysis approach in combination with the bootstrap procedure (i.e., ITCe) to address different trial counts, our previous finding demonstrated significantly enhanced ITCe values when participants could consciously recognize the facial identity of the target stimulus (FI trials: the mean number of trials $\pm$ SD = 55 ± 6) compared with those when they could not see the target (FM trials: 96 ± 14; cluster-based permutation test,

Fig. 4. Phase-clustering differences between FI and FM trials as indicated by (A) the benchmark finding, (B) CS, (C) ITC and (D) ITCz measures. The black asterisks in the scalp topographies highlight a cluster of sensors exhibiting significant differences at (A) 10 Hz:390 ms, (B) 10 Hz:390 ms, (C) 47 Hz:660 ms and (D) 20 Hz:490 ms. In the time-frequency representations of the representative sensors, the black lines indicate the onset of the first mask, and the dotted lines indicate the onset of the stimulus. The brackets highlight the time-frequency window exhibiting significant differences.

$p < 0.001$). This significant effect occurred at the alpha (10 Hz) frequency in the right frontal-parietal-temporal sites between 240 and 580 ms after the onset of the first mask (Fig. 4A). In Study III, we used this previous finding as a benchmark to evaluate the extent to which the ITC, ITCz and CS measures could properly replicate the previous finding.

After computing the CS, ITCz or ITC values on every sensor-time-frequency data point, cluster-based permutation tests were conducted to compare the values between FI and FM trials. Comparable to the benchmark finding, the CS measure revealed higher phase clustering in FI trials than that in FM trials ($p = 0.03$, Fig. 4B); this finding was similarly observed at 10–12 Hz in the right frontal-parietal-temporal sites between 250 and 520 ms after the onset of the first mask. Although phase clustering, as indexed by the classic ITC measure, was also enhanced in FI trials, the effect was widely spread over most of the sensor-time-frequency window ($p < 0.001$, Fig. 4C). In contrast to the CS and ITC measures, the ITCz measure showed higher phase clustering in FM trials than that in FI trials. This opposite observation was found at 12–28 Hz between 370 and 580 ms after the onset of the first mask and was mainly located over the occipital sensors but extended to the right temporal and left central-parietal sensors ($p = 0.005$, Fig. 4D).

To illustrate the obtained results relative to the benchmark finding, we respectively calculated two indices, the sensitivity and specificity, of each measure (Fig. 5A and B). Here, sensitivity denotes to what extent each measure could "accurately" detect the significant results as indicated by the benchmark finding, whereas specificity denotes to what extent each measure could "accurately" detect the non-significant results as indicated by the benchmark finding (see Section 2.3.3 for details). On the one hand, the ITC measure exhibited the greatest sensitivity, followed by the CS measure, with zero sensitivity for the ITCz measure. On the other hand, the CS measure exhibited the greatest specificity, followed by the ITCz measure, with the lowest specificity for the ITC measure. In sum, when the previous finding was used as a benchmark, the CS measure outperformed the ITCz and ITC measures and provided the most satisfactory results in terms of sensitivity and specificity.

ROC analysis was further conducted to better evaluate the validity of each phase clustering measure (see Section 2.3.3 for details). In previous cluster-based permutation tests, the criterion that defined cluster selection was fixed (i.e., a minimum number of 2 neighborhood sensors). For this analysis, we manipulated this criterion so that ROC curves could be constructed and AUC values were subsequently calculated. In support of a previous report, the AUC values indicated that CS (AUC = 0.801) effectively predicted the benchmark finding, whereas ITC (0.5012) and ITCz (0.5021) only exhibited chance performance in prediction (Fig. 5C).

4. Discussion

To "correct" for the problem of sample size effects in the phase-clustering measurement of a single brain signal, this study introduced CS as a new measure. In addition, we systematically examined the previous "correction" measure, ITCz (Bonnefond and Jensen, 2012; Cohen, 2014; Samaha et al., 2015), and the classic measure, ITC (Tallon-Baudry et al., 1996). After comparing their performances in a series of studies using either artificial or real datasets, the overall results suggest that CS may serve as an optimal measure to quantify phase clustering.

The current study first provides a novel insight into the nature of the previous "correction" measure, ITCz. Similar to the classic ITC measure, ITCz values vary as a function of sample size, as shown in Study I. Moreover, in most of the parametric conditions, the sample-size effects persist even when a reasonable number of sample sizes are accumulated. However, the general patterns of the sample-size effects differ between ITCz and ITC. On the one hand, consistent with prior literature (Edwards et al., 2009), ITC tends to overestimate the degree of phase clustering with small sample sizes, but the exact degree of a sample-size bias depends on the nature of the distribution (i.e., distribution concentration) that artificial data are sampled from and the frequency ranges in which real data are located. On the other hand, ITCz overestimates the degree of phase clustering with larger sample sizes for artificial datasets, whereas for actual datasets, the sample-size bias varies according to fre-

Fig. 5. (A) Sensitivity, (B) specificity and (C) the AUC of the phase-clustering measures. For the results obtained from the cluster-based permutation tests with a fixed cluster selection criterion, sensitivity denotes to what extent each measure could "accurately" detect the significant results as indicated by the benchmark finding, whereas specificity denotes to what extent each measure could "accurately" detect the non-significant results as indicated by the benchmark finding. The AUC was obtained by varying the criterion that defined cluster selection in the cluster-based permutation tests. The AUC values denote how well each measure could predict the benchmark finding, with a chance level of 0.5 and a maximal performance of 1.

quency ranges, but there is no systematic relationship between the two.

To further explore the validity of ITCz compared with that of ITC, we additionally evaluated how ITCz and ITC performed in situations where two artificial (Study II) or two actual datasets (Study III) had known or assumed real effects in phase clustering. The overall results agree with the phenomena observed in Study I. First, for the ITC measure, because of serious overestimation with small sample sizes, the results in Study II show that the ITC measure misidentified datasets with smaller sample sizes (i.e., "data20_low" and "data20_same") as having stronger phase clustering and thereby could not accurately assess phase-clustering differences between the datasets. In Study III, because there were fewer FI trials than FM trials (Hsu and Yang, 2017 in press), when FI trials are compared with FM trials, ITC overestimates the degree of phase clustering in FI trials and thereby produces a large number of false positive effects outside the benchmark finding, leading to low specificity.

Importantly, the ITCz measure also appears to be an unreliable measure even though it has been previously used to address sample-size bias. Although ITCz is able to "correct" the bias present in ITC and in turn differentiate two datasets with different degrees of phase clustering and different sample sizes (Study II-1), its ability to detect two datasets with the same degree of phase clustering but different sample sizes is unsatisfactory (Study II-2). This unsatisfactory outcome could be attributed to overestimation of the degree of phase clustering in artificial datasets with large sample sizes by ITCz. For real datasets, because the ITCz measure produces a mixed bias in assessing phase clustering between 12 and 28 Hz as revealed by Study I, this complex pattern of effects yields a significant consequence. As evident in Study III, in contrast to the benchmark and all other findings, the result direction is completely reversed (i.e., stronger phase clustering in FM trials instead of in FI trials). Moreover, all the significant sensor-time-frequency data points identified by ITCz fall outside the benchmark finding, indicating the low sensitivity of this measure.

Compared with ITC and ITCz, CS is robust against the sample size effect, and the values converge to a steady state even with small sample sizes (Study I). Moreover, several different aspects of examinations (Study II and III) further confirm that CS may successfully detect the true nature of phase clustering between datasets with differing sample sizes. However, the CS measure also exhibits some limitations. First, CS is not completely free from sample-size effects. Study I indicates that the CS values become unstable with extremely small sample sizes. We suggest that this phenomenon could be attributed to the fact that when a sample size is extremely small, samples located at the extremes of a given population spectrum could be occasionally selected, such as samples with a completely opposite or identical angle. This situation may consequently result in highly negative or positive CS, leading to measurement instability. Accordingly, we parsimoniously suggest employing the CS measure when there are at least 20 samples for analysis. Second, Study III shows that although CS exhibits stronger specificity in detecting non-significant benchmark results, its sensitivity in detecting significant benchmark results is imperfect. These observations suggest that CS has less power in detecting real effects compared with ITCe when a sufficient number of trials remains after trial reduction in ITCe. Alternatively, CS may be regarded as a better measure, as the same finding can be interpreted

as CS produces less Type I errors than the ITCe approach. A future investigation is needed to thoroughly examine this issue.

In addition to being sample size free, the CS measure demonstrates several other appeals. First, CS has an intuitive interpretation of phase clustering, as it indicates the similarity of phases within a neurophysiological signal in terms of the mean cosine angle of all given pairs of phase vectors. In other words, if there is a "preferred" phase angle within a signal, the distribution of pairwise differences of phase vectors will be more strongly clustered around an average value, leading to high CS. By contrast, when no "preferred" phase angle is present, the distribution of pairwise differences of phase vectors will be randomly distributed around the phase circle, leading to low CS. In addition to the interpretation issue, CS appeals to empirical studies. Because CS is able to reach stability even with low sample size, this measure is especially applicable in some experimental situations wherein data are scarce. To conclude, the current study not only reveals the incompetence of the ITCz and ITC measures but also provides converging evidence showing the strengths of the CS measure. Therefore, we suggest including the CS measure in future explorations of phase clustering due to its validity and practicality.

Acknowledgements

This work was supported by the Ministry of Science and Technology of Taiwan, R.O.C. (MOST 104-2410-H-004-049).

References

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. ACM press, New York.
Berens, P., 2009. CircStat: a MATLAB toolbox for circular statistics. J. Stat. Software 31, 1–21.
Bonnefond, M., Jensen, O., 2012. Alpha oscillations serve to protect working memory maintenance against anticipated distracters. Curr. Biol. 22, 1969–1974.
Cohen, M.X., 2014. Analyzing Neural Time Series Data: Theory and Practice. The MIT Press, Cambridge, Massachusetts.
Doelling, K.B., Poeppel, D., 2015. Cortical entrainment to music and its modulation by expertise. Proc. Natl. Acad. Sci. USA 112, E6233–E6242.
Drewes, J., van Rullen, R., 2011. This Is the Rhythm of Your Eyes: the phase of ongoing electroencephalogram oscillations modulates saccadic reaction time. J. Neurosci. 31, 4698–4708.
Edwards, E., Soltani, M., Kim, W., Dalal, S.S., Nagarajan, S.S., Berger, M.S., Knight, R.T., 2009. Comparison of time-frequency responses and the event-related potential to auditory speech stimuli in human cortex. J. Neurophysiol. 102, 377–386.
Fisher, N.I., 1993. Statistical Analysis of Circular Data. Cambridge University Press, Cambridge England; New York, NY, USA.
Grandchamp, R., Delorme, A., 2011. Single-trial normalization for event-related spectral decomposition reduces sensitivity to noisy trials. Front. Psychol. 2.
Hotho, A., Nürnberger, A., Paaß, G., 2005. A brief survey of text mining. Ldv Forum 20, 19–62.
Hsu, S.-M., Yang, Y.-F., 2017. Temporal neural mechanisms underlying conscious access to different levels of facial stimulus contents. J. Neurophysiol., in press.
Jeffreys, H., 1961. The Theory of Probability, 3rd edition. Oxford University Press, Oxford.
Kutil, R., 2012. Biased and unbiased estimation of the circular mean resultant length and its variance. Statistics 46, 549–561.
Lachaux, J.P., Rodriguez, E., Martinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. Hum. Brain Mapp. 8, 194–208.
Lakatos,