Choosing clustering method

There is no definitive answer to your question, as even within the same method the choice of the distance to represent individuals (dis)similarity may yield different result, e.g. when using euclidean vs. squared euclidean in hierarchical clustering. As an other example, for binary data, you can choose the Jaccard index as a measure of similarity and proceed with classical hierarchical clustering; but there are alternative approaches, like the Mona (Monothetic Analysis) algorithm which only considers one variable at a time, while other hierarchical approaches (e.g. classical HC, Agnes, Diana) use all variables at each step. The k-means approach has been extended in various way, including partitioning around medoids (PAM) or representative objects rather than centroids (Kaufman and Rousseuw, 1990), or fuzzy clustering (Chung and Lee, 1992). For instance, the main difference between the k-means and PAM is that PAM minimizes a sum of dissimilarities rather than a sum of squared euclidean distances; fuzzy clustering allows to consider "partial membership" (we associate to each observation a weight reflecting class membership). And for methods relying on a probabilistic framework, or so-called model-based clustering (or latent profile analysis for the psychometricians), there is a great package: Mclust. So definitively, you need to consider how to define the resemblance of individuals as well as the method for linking individuals together (recursive or iterative clustering, strict or fuzzy class membership, unsupervised or semi-supervised approach, etc.).

Usually, to assess cluster stability, it is interesting to compare several algorithm which basically "share" some similarity (e.g. k-means and hierarchical clustering, because euclidean distance work for both). For assessing the concordance between two cluster solutions, some pointers were suggested in response to this question, Where to cut a dendrogram? (see also the cross-references for other link on this website). If you are using R, you will see that several packages are already available in Task View on Cluster Analysis, and several packages include vignettes that explain specific methods or provide case studies.

Cluster Analysis: Basic Concepts and Algorithms provides a good overview of several techniques used in Cluster Analysis. As for a good recent book with R illustrations, I would recommend chapter 12 of Izenman, *Modern Multivariate Statistical Techniques* (Springer, 2008). A couple of other standard references is given below:

- Cormack, R., 1971. A review of classification. *Journal of the Royal Statistical Society, A* 134, 321–367.
- Everitt, B., 1974. *Cluster analysis*. London: Heinemann Educ. Books.
- Gordon, A., 1987. A review of hierarchical classification. *Journal of the Royal Statistical Society, A* 150, 119–137.
- Gordon, A., 1999. *Classification*, 2nd Edition. Chapman and Hall.
- Kaufman, L., Rousseuw, P., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, Wiley.

What is a meaning of "p-value F" from Friedman test?

I generally used `friedman.test()` which doesn't return any F statistic. If you consider that you have b blocks, for which you assigned ranks to observations belonging to each of them, and that you sum these ranks for each of your a groups (let denote them sum R_i), then the Friedman statistic is defined as

$$F_r = \frac{12}{ba(a+1)}\sum_{i=1}^{a} R_i^2 - 3b(a+1)$$

and follows a $\chi^2(a-1)$, for a and b sufficiently large. Quoting Zar (*Biostatistical Analysis*, 4th ed., pp. 263–264), this approximation is conservative (hence, test has low power) and we can use an F-test, with

$$F_{\text{obs}} = \frac{(b-1)F_r}{b(a-1)-F_r}$$

which is to be compared to an F distribution with $a-1$ and $(a-1)(b-1)$ degrees of freedom.

The best measure of reliability for interval data between 0 and 1

Referring to your comments to @Henrik, I'm inclined to think that you rather have continuous measurements on a set of objects (here, your similarity measure) for 6 raters. You can compute an intraclass correlation coefficient, as described here Reliability in Elicitation Exercise. It will provide you with a measure of agreement (or concordance) between all 6 judges wrt. assessments they made, or more precisely the part of variance that is explained by between-rater variance. There's a working R script in appendix.

Note that this assumes that your measures are considered as real valued measurement (I refer to @onestop's comment), not really proportions of similarity or whatever between your paired sounds. I don't know of a specific version of the ICC for % or values bounded on an interval, only for binary or ranked data.

Update:

Following your comments about parameters of interest and language issue:

- There are many other online ressources on the ICC; I think David Howell provides a gentle and well illustrated introduction to it. Its discussion generalize to k-sample (judges/raters) without any difficulty I think, or see this chapter from Sea and Fortna on Psychometric Methods. What you have to think to is mainly whether you want to consider your raters as an unique set of observers, not necessarily representative of all the raters that would have assess your object of measurement (this is called a fixed effect), or as a random sample of raters sampled from a larger (hypothetical) population of potential raters: in the former case, this corresponds to a one-way anova or a consistency ICC, in the latter case we talk about an agreement ICC.

- A colleague of mine successfully used Kevin Brownhill's script (from Matlab Central file exchange). The ICC you are interested in is then `cse=3` (if you consider that your raters are not representative of a more general population of raters).

How do you draw structural equation/MPLUS models?

I use the psych R package for CFA and John Fox's sem package with simple SEM. Note that the graphical backend is graphviz. I don't remember if the lavaan package provides similar or better facilities.

Otherwise, the Mx software for genetic modeling features a graphical interface in its Windows flavour, and you can export the model with path coefficients.

What graphical techniques are used in Structural Equation Modeling?

I worked with Laura Trinchera who contributed a nice R package for PLS-path modeling, plspm. It includes several graphical output for various kind of 2- and k-block data structures.

I just discovered the plotSEMM R package. It's more related to your second point, though, and is restricted to graphing bivariate relationships.

As for recent references on diagnostic plot for SEMs, here are two papers that may be interesting (for the second one, I just browsed the abstract recently but cannot find an ungated version):

1. Sanchez BN, Houseman EA, and Ryan LM. Residual-Based Diagnostics for Structural Equation Models. *Biometrics* (2009) 65, 104–115

2. Yuan KH and Hayashi K. Fitting data to model: Structural equation modeling diagnosis using two scatter plots, *Psychological Methods* (2010)

Data transformation for Principal Components Analysis from different likert scales

As suggested by @whuber, you can "abstract" the scale effect by working with a standardized version of your data. If you're willing to accept that an interval scale is the support of each of your item (i.e. the distance between every two response categories would have the same meaning for every respondents), then linear correlations are fine. But you can also compute polychoric correlation to better account for the discretization of a latent variable (see the R package polycor). Of note, it's a largely more computer-intensive job, but it works quite well in R.

Another possibility is to combine optimal scaling within your PCA, as implemented in the homals package. The idea is to find a suitable non-linear transformation of each scale, and this is very nicely described by Jan de Leeuw in the accompagnying vignette or the JSS article, Gifi Methods for Optimal Scaling in R: The Package homals. There are several examples included.

For a more thorough understanding of this approach with any factorial method, see the work of Yoshio Takane in the 80s.

Similar points were raised by @Jeromy and @mbq on related questions, Does it ever make sense to treat categorical data as continuous?, How can I use optimal scaling to scale an ordinal categorical variable?

How does one calculate Cohen's d and confidence intervals after logit in Stata?

Cohen's d is not directly available in Stata, and you have to resort on external macros, e.g. sizefx (`ssc install sizefx`). It works fine if you have to series of values, but I found it less handy when you work with a full data set because there's no possibility to pass options to this command (e.g. `by()`).

Anyway, you can still use the original formula (with pooled SDs),

$$\delta_c = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

where $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}$.

Here is an example by hand:

```
. webuse lbw
. logit low age smoke
. graph box age, by(low)
. tabstat age, by(low) statistics(mean sd N)

Summary for variables: age
     by categories of: low (birthweight<2500g)

     low |      mean        sd         N
---------+------------------------------
       0 |  23.66154  5.584522       130
       1 |  22.30508  4.511496        59
---------+------------------------------
   Total |   23.2381  5.298678       189
----------------------------------------

. display "Cohen's d: = " (23.66154-22.30508) / sqrt((129*(5.584522)^2+58*(4.51

Cohen's d: = .25714326
```

This is in agreement with what R would give:

```
library(MBESS)
res <- smd(Mean.1=23.66154, Mean.2=22.30508,
           s.1=5.584522, s.2=4.511496, n.1=130, n.2=59)
ci.smd(smd=res, n.1=130, n.2=59, conf.level=0.95)
```

that is an effect size of 0.257 with 95% CI [–0.052;0.566].

In contrast, `sizefx` gives results that differ a little (I have use `separate age, by(low)` and collapse the results in a new data window, here two columns labeled `age0` and `age1`), the ES version calculated above corresponding to what is referred to as Hedge's g below (unless I miss something in the code I read):

```
. sizefx age0 age1

Cohen's d and Hedges' g for: age0 vs. age1
Cohen's d statistic (pooled variance) = .26721576
Hedges' g statistic = .26494154

Effect size correlation (r) for: age0 vs. age1
ES correlation r = .13243109
```

Inter-rater reliability between similarity matrices

My first idea would be to try some kind of cluster analysis (e.g. hierarchical clustering) on each similarity matrix, and compare the classification trees across raters. We can derive a similarity index from all dendrograms, as discussed here, A measure to describe the distribution of a dendrogram, or in this review, Comparing Clusterings - An Overview from Wagner and Wagner.

You benefit from working with already existing distance matrices, thus such methods will really reflect the nature of your data, and you can still derive a single numerical value to quantify the closeness of method-specific assessments. The following article may be interesting, if you need to refer to existing work:

> Hamer, RM and Cunningham, JW. Cluster Analyzing Profile Data Confounded with Interrater Differences: A Comparison of Profile Association Measures. *Applied Psychological Measurement* (1981) 5(1): 63–72.

Another approach would be to apply some kind of Principal Component Analysis on each similarity matrix, and keep only the first principal component (the linear combination of all 100 items that account for the maximum of variance). More precisely, as you work with (dis)similarity indices or a particular distance/proximity metric, it is sometimes referred to as Principal Coordinate Analysis or Multidimensional Scaling (MDS), although PCA and MDS would yield similar results when dissimilarities are defined as euclidean distances. There is a working example in Izenman's book (*Modern Multivariate Statistical Techniques*, chapter 13, "perceptions of color in human vision", pp. 468–470) and a discussion on so-called *all-pairs design* pp. 471–472. You can then compare the 6 linear combinations (i.e., the weights associated to each sound by rater-specific MDS) to assess their consistency across raters. There, an ICC (as described in my previous answer) could make sense, but I don't know of any application of it in this particular case.

How to create a barplot diagram where bars are side-by-side in R

I shall assume that you are able to import your data in R with `read.table()` or the shorthand `read.csv()` functions. Then you can apply any summary functions you want, for instance `table` or `mean`, as below:

```
x <- replicate(4, rnorm(100))
apply(x, 2, mean)
```

or

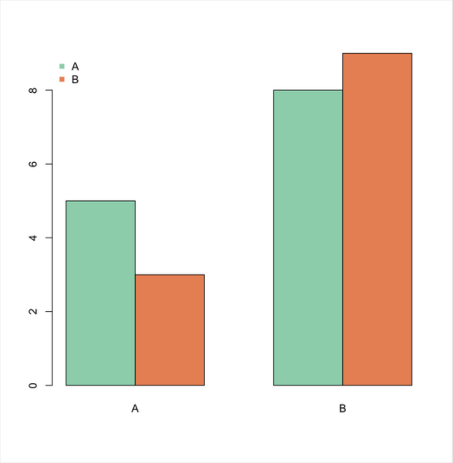
```
x <- replicate(2, sample(letters[1:2], 100, rep=T))
apply(x, 2, table)
```

The idea is to end up with a matrix or table for the summary values you want to display.

For the graphical output, look at the `barplot()` function with the option `beside=TRUE`, e.g.

```
barplot(matrix(c(5,3,8,9),nr=2), beside=T,
        col=c("aquamarine3","coral"),
        names.arg=LETTERS[1:2])
legend("topleft", c("A","B"), pch=15,
       col=c("aquamarine3","coral"),
       bty="n")
```

The `space` argument can be used to add an extra space between juxtaposed bars.



Comparing test-retest reliabilities

Both situations are specific cases of test-retest, except that the recall period is null in the first case you described. I would also expect a larger agreement in the former case, but that may be confounded with a learning or memory effect. A chance-corrected measure of agreement, like Cohen's kappa, can be used with binary variables, and bootstraped confidence intervals might be compared in the two situations (this is better than using κ sampling variance directly). This should give an indication of the reliability of your measures, or in this case diagnostic agreement, at the two occasions. A McNemar test which tests for marginal homogeneity in matched pairs can also be used.

An approach based on the intraclass correlation is still valid and, provided your prevalence is not extreme, should be closed to

- a simple Pearson correlation (which, for binary data, is also called a phi coefficient) or the tetrachoric version suggested by @Skrikant,
- the aforementioned kappa (for a large sample, and assuming that the marginal distributions for case at the two occasions are the same, $\kappa \approx$ ICC from a one-way ANOVA).

About your bonus question, you generally need 3 time points to separate the lack of (temporal) stability — which can occur if the latent class or trait your are measuring evolve over time — from the lack of reliability (see for an illustration the model proposed by Wiley and Wiley, 1970, *American Sociological Review* 35).