

CRITERION SETTING FOR OBJECTIVE, FOURIER ANALYSIS BASED PILOT PERFORMANCE METRICS

Nicholas R. Johnson, Esa M. Rantanen, and Donald A. Talleur
Aviation Human Factors Division
University of Illinois at Urbana-Champaign

This study reports the development and evaluation of time series based objective pilot performance metrics. From a previously developed array of autocorrelation and Fourier analysis based metrics, five Fourier-based metrics that employed a threshold value were chosen to investigate their effectiveness in separating pilots who, based on instructor pilot (IP) evaluations, had either passed or failed a particular segment of an instrument proficiency check flight. An instrument landing system (ILS) approach was chosen for analysis based on IP feedback of what flight segments were most difficult to evaluate, had greatest sensitivity to overall pilot performance, and greatest criticality to the flying task. Further analysis showed that an optimal value for the criterion value could be found that most effectively separated those pilots that had passed the ILS segment from those who had failed. Criterion setting methods without external criteria using multidimensional scaling and cluster analysis techniques are also discussed.

INTRODUCTION

Measurement is a prerequisite for all scientific research and subsequent engineering applications (Chapanis, 1959). In aviation research, pilot performance is often the response variable of greatest importance, on which the impact of various manipulations in training program design (e.g., Taylor et al., 2002) or methods of skill maintenance (e.g., Talleur et al., 2003) is examined. However, pilot performance is a complex and multifaceted construct, and its adequate measurement as a dependent variable for research purposes is consequently problematic. The most common method of pilot performance measurement is through direct observation by a qualified instructor pilot (IP). If a standardized checklist is used where all the items a subject is to be evaluated on are explicitly defined and the IP evaluators are adequately trained to achieve reasonable inter-rater and intra-rater reliability, this method can effectively capture much of the “whole” of the subject pilot’s performance. However, an IP may not be able to provide sufficiently accurate quantitative data for research purposes, due to the limitations of human observation capabilities. Likewise, the IP may not be able to record observed performance data at a sufficient frequency to study the variable in question. This is particularly the case in observation of simultaneous events. For these reasons, objective pilot performance measures are highly desirable.

Objective Pilot Performance Metrics

Rantanen and Talleur (2001) evaluated a battery of 5 specific pilot performance metrics that were derived from flight data recorder data for 9 different flight parameters from a number of different instrument flight maneuvers. It was shown that many of these metrics closely corresponded to subjective IP judgments on pilot performance. However, as these metrics were practically static and as such did not measure the behavior of the flight parameters over time, they have recently been being complemented with additional metrics derived from time series analyses of the same parameters (Johnson, Rantanen, & Talleur, in press). These

new metrics utilize spectral (Fourier) and autocorrelation analyses. Such analyses have been identified as being potentially beneficial to pilot performance measurement when combined with the more standard amplitude-distribution measures like mean, standard deviation and root mean square error (Semple, Cotton & Sullivan, 1981). Indeed, Hills and Eddowes (1975) and Vreuls et al. (1975) used spectral analysis to derive measures that were used in multi-variate analysis to discriminate between levels of pilot performance in training.

Using Fourier analysis, a time series of data can be decomposed into spectral or frequency components. This decomposition allows for an explicit representation of the underlying frequencies occurring in the time series (Figure 1). A measure of the magnitude of each frequency component is called the *Power Spectral Density* (PSD) and represents the weighted contribution each frequency component makes to the original time series of data. A plot of PSD against frequency is termed a periodogram.

Our guiding hypothesis used to develop the Fourier analysis based metrics was that better pilots would exhibit a larger range of frequencies of aircraft control than less able pilots, who may only control the aircraft with lower frequency control inputs. Under this hypothesis, the periodogram from a good pilot’s time series of data would contain a greater range of spectral components with significant PSD values than a corresponding periodogram from poor pilot’s data. Conversely, we would expect the periodogram from a poor pilot’s time series of data to contain a greater proportion of low frequency spectral components than a good pilot’s.

Figure 1 shows periodograms derived from data of two pilots’ vertical speed time series recorded during a straight and level segment of a research flight. The pilots either passed or failed the flight segment based on a subjective IP evaluation. Qualitative differences between the passed and failed pilots’ periodograms can be seen that are consistent with the hypothesis that less skilled pilots will show a greater proportion of low frequency control inputs. While the failed pilot’s periodogram only contains significant frequency components at frequencies less than approximately 0.05Hz,

the pass pilot's periodogram contains a greater range of frequencies that contribute significantly to the original time series. The metrics that were developed with Fourier methods are used to quantify both the range and magnitude of these significant frequency components.

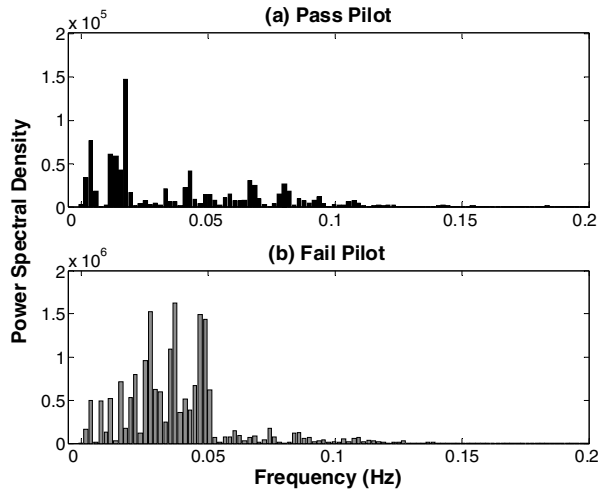


Figure 1. Fourier decomposition of two pilots' performance on vertical speed in a level-flight segment.

In determining what spectral components of the Fourier decomposition were significant (i.e., sufficiently large to reflect different performances of pilots), a criterion PSD value, v_c was set. Consequently, 7 Fourier-analysis based metrics were developed: (1) PSD mean and (2) PSD standard deviation (SD) of the spectral components, (3) the number of spectral components with PSD greater than v_c , (4) the mean and (5) SD of spectral components with PSD greater than v_c , and (6) the mean frequency and (7) SD of the frequencies of spectral components with PSD greater than v_c (Johnson, Rantanen, & Talleur, in press). Because the data ranges of the time series vary greatly between flight parameters and individual pilots, spectral magnitudes in the Fourier decomposition will also vary greatly between parameters and pilots. Thus setting a single critical value to be used across all pilots' flight parameters will not achieve the desired level of sensitivity. Therefore, a relative v_c that was set as a fraction of the maximum value of the spectral components was used in the analysis of the latter five metrics (3–7). This approach will also allow manipulation of v_c in order to find the value that produces maximum sensitivity in distinguishing between good and poor pilots.

Purpose of the Study

An earlier study (Johnson, Rantanen, & Talleur, in press) showed the utility of the Fourier analysis based metrics in differentiating between a good pilot and a poor pilot, judged against IP evaluation of Instrument Proficiency Check (IPC) flight segments and segment components (e.g., tracking accuracy of a glide slope). The goal of the present study was to determine how the criterion value could be set in a way that optimized the sensitivity to subjective pilot performance evaluations for a larger number of pilots. That is, we changed

the criterion value progressively over an appropriate range in order to maximize the separation of the two pilot groups, those who passed and those who failed an IPC flight, as judged by an IP. In effect, this task is a form of cluster analysis.

METHOD

Data was collected from IPC flights in an aircraft equipped with a flight data recorder that measured airspeed, altitude, vertical speed, heading, pitch, roll, ball deflection, course deflection indications (CDI) and glide slope indications (GS) (Rantanen & Talleur, 2001). Each flight parameter was sampled at 1 Hz and data were stored on an on-board computer. An IPC flight consists of 14 distinct segments, including VHF Omnidirectional Range (VOR) tracking, Instrument Landing System (ILS) approach, non-precision VOR approach, holding pattern, and steep turns. Data from the flights were segmented using a data visualization tool (see Rantanen & Talleur, 2001 for a description of the tool and data preprocessing procedure) before further analyses.

To determine the particular instrument flight maneuvers that would most benefit from objective performance metrics, a survey of approximately 50 certified flight instructors, instrument (CFII) at the Institute of Aviation of the University of Illinois at Urbana-Champaign was conducted. The respondents were asked to rate the maneuver components of the IPC flight each along three scales: (1) The difficulty of observation and recording of performance on the particular element, (2) the criticality of the particular element in terms of the overall success of the student pilot, and (3) the sensitivity of the particular element in bringing out differences in the performance of pilots with different skill level. A weighted average of the rankings was used to pinpoint maneuvers and elements (e.g., altitude control during procedure turn) most challenging to the IP and that hence should be augmented by objective performance metrics.

The results of the survey showed that instructors consistently ranked ILS glide slope (GS) and localizer (CDI) tracking at or near the top of the three scales. ILS glide slope tracking ranked second for difficulty and criticality and first for sensitivity out of the 48 flight elements monitored during the IPC flight. ILS localizer tracking ranked fourth for difficulty, third for criticality and second for sensitivity. Based on these results, the ILS GS and localizer CDI elements were chosen for the present analysis.

Time series analyses were carried out in MATLAB version 6.5.1. For criterion setting, the value of v_c was incrementally adjusted from zero to 95% of the maximum value of the magnitude of the spectral components (see Figure 2) in 5% steps. At the upper limit, there were often only one or two spectral components with magnitude greater than the mean of all components (i.e., a sharply peaked distribution). The resultant Fourier metrics from each iteration of v_c were subsequently analyzed by one-way Analysis of Variance (ANOVA) to predict metric value from group (pass or fail) membership. The v_c value that maximized the F -value was then chosen. Alternatively, the magnitude of the difference in group means could be used in place of F -values.

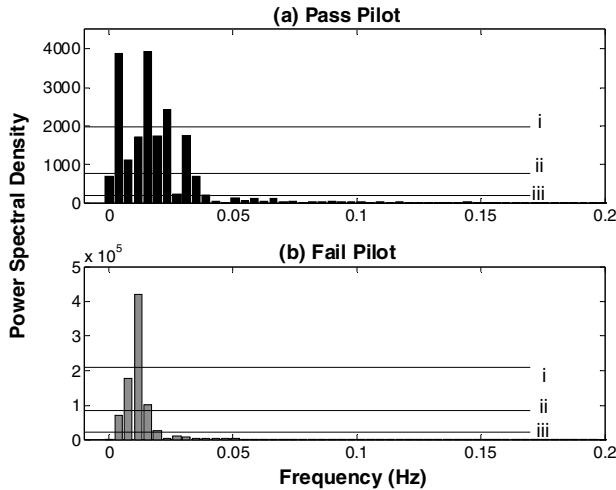


Figure 2. Periodograms of pilots who passed (a) and failed (b) the CDI evaluation, with criterion values superimposed. Note the different y-axis scales. As the criterion value, v_c , is lowered from (i) 50% to (ii) 20% to (iii) 5% of the maximum value of the PSD, more components are counted and used in the metrics described in the text.

RESULTS

Twenty-three time series of flight data were used for analysis. Due to a missing IP evaluation of the localizer CDI performance for one pilot, only 22 time series were used in CDI analysis. From the total of 10 Fourier metrics used, 4 produced ANOVA results that were successful in separating the “pass” pilots from the “fail” pilots for the particular flight element (CDI or GS) at the $\alpha < .05$ significance level. That is, the means of the metric values for the two groups differed significantly. The mean of the spectral components with PSD exceeding v_c for the CDI produced significant results for most iterations, $F(1,20) > 4.35$, $p < .05$. The same measure for the GS element produced significant differences across all iterations of v_c , $F(1,21) > 4.32$, $p < .05$. The SD of spectral component magnitudes only produced significant differences between pilot groups from iteration 1 ($v_c=0$) to 10 ($v_c=45\%$ of maximum), with the exception of iteration 3 ($v_c=10\%$ of maximum) for the CDI element and from iteration 1 to iteration 9 ($v_c=40\%$ of maximum) for the GS element. These results are presented in Figures 3 and 4. The other metrics did not show significant differences between the pilot groups.

From the results (Figures 3 and 4) we see that setting the criterion value close to zero for the mean of the spectral components will produce the highest F-values (or alternatively, largest differences in means) for separation of the two pilot groups for the CDI element. In addition setting v_c at a relatively high level will also produce a good separation. This pattern is again repeated for GS but the F-values are consistently higher, giving greater flexibility in setting v_c . For SD of the spectral components, it is necessary to set v_c at relatively low values: between 15% and 50% of maximum for CDI and between 0 and 40% for GS. It should also be noted that when v_c is set to zero, metrics 4 and 5 reduce to metrics 1

and 2, the PSD mean and standard deviation of all spectral components in the Fourier decomposition.

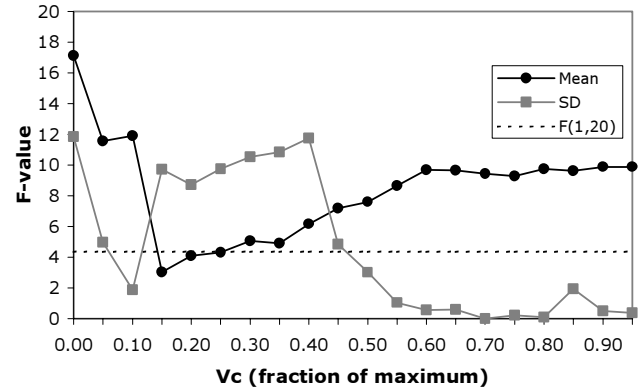


Figure 3: Mean and SD of spectral components with PSD greater than v_c from CDI tracking performance. The value of the F-distribution that signifies significance at the 0.05 level is $F(1,20) = 4.35$.

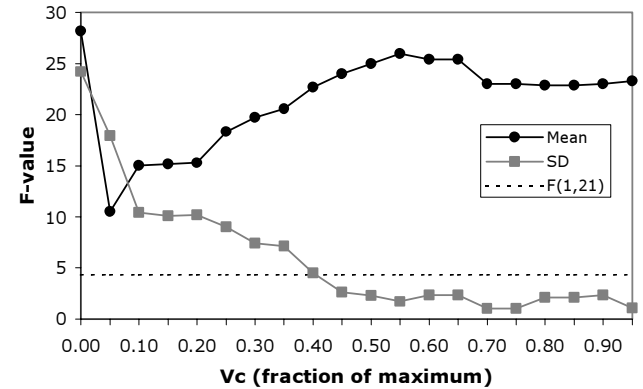


Figure 4: Mean and SD of spectral components with PSD greater than v_c from GS tracking performance. The value of the F-distribution that signifies significance at the 0.05 level is $F(1,21) = 4.32$.

One reason for the decrease in F-values with iteration number for the standard deviation of spectral component magnitudes in both CDI and GS elements was the fact that as v_c increased towards the maximum value of the spectral component magnitudes, there would be a point where only one component (the largest) was being counted and used for further analysis. While this was not a problem for the mean it is obviously a problem for SD. When the v_c had reached the level where only one component met that criterion value, SD would be zero for the remaining iterations and consequently reduce the effectiveness of the metric in distinguishing pilot group means. To get around this issue, the criterion value should be set so that at least two spectral components are being counted and used in the subsequent SD calculations.

While the 4 metrics described above produced statistically significant differences in the means based on pass or fail group membership, the metrics did not completely distinguish

between the two groups. That is, there was some overlap in metric values between the two groups (Figures 5 and 6). Ideally, the metrics should completely separate the pass and fail pilot groups

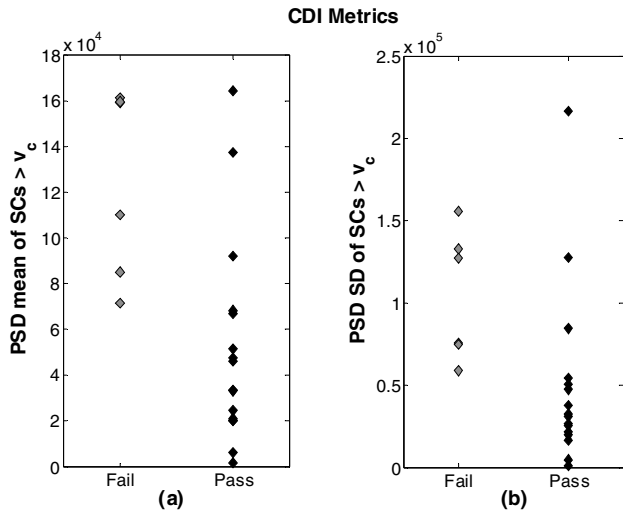


Figure 5. CDI Metric value distributions for pass and fail groups when $v_c = 10\%$. (a) Metric 3: Mean of spectral components (SCs) with PSD greater than v_c ; (b) Metric 4: SD of SCs with PSD greater than v_c .

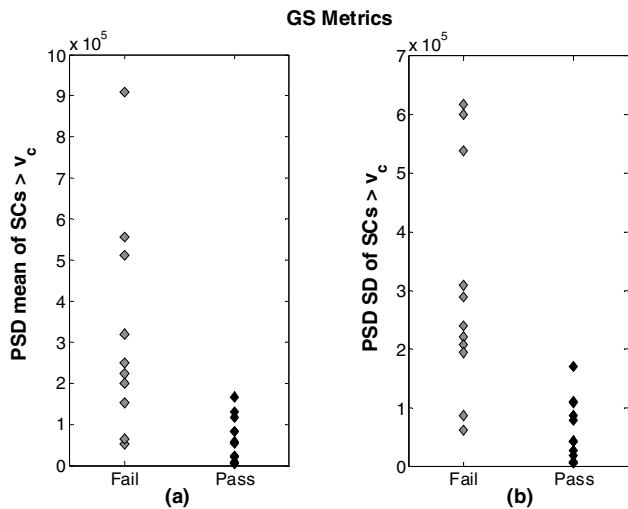


Figure 6. GS Metric value distributions for pass and fail groups when $v_c = 10\%$. (a) Metric 3: Mean of spectral components (SCs) with PSD greater than v_c ; (b) Metric 4: SD of SCs with PSD greater than v_c .

DISCUSSION

These results show that of the 5 Fourier based metrics the PSD mean and standard deviation of spectral components closely match the pass and fail evaluations by IPs. The results also show that an optimum value of v_c can be chosen to

provide best sensitivity in distinguishing between the two pilot groups.

When the criterion value was set to zero, the criterion based metrics reduced to the over all metrics where all spectral components in the Fourier decomposition were counted for the mean and SD calculations. From Figures 3 and 4 it is apparent that the mean and SD metrics based on all spectral components showed as good or greater sensitivity to group than the metrics using a non-zero criterion value. That is, F-values for iterations where $5\% \leq v_c \leq 95\%$ were less than or equal to the F-values when $v_c = 0$. This would suggest that simply using all spectral components in the metrics will produce the best sensitivity, without having to implement additional computations with criterion values. Further analysis involving other flight parameters may reveal whether this result is true in general. In either case, the iterative approach to metric generation may be useful when combining additional metrics.

Given that the individual Fourier based metrics did not completely separate the two pilot groups along one axis (Figures 5 and 6), exploring the combination of two closely related metrics in a two-axis situation would be an interesting extension of this work. This method will allow setting of optimum threshold values for both metrics to maximize separation between well and poorly performing pilots without external criteria (i.e., IP evaluations).

References

- Chapanis, A. (1959). *Research techniques in human factors*. Baltimore, MD: Johns Hopkins University Press.
- Johnson, N. R., Rantanen, E. M., & Talleur, D. A. (in press). Time series based objective pilot performance measures. *International Journal of Applied Aviation Studies (IJAAS)*.
- Hills, J.W. & Eddowes, E.E. (1974). *Further Development of Automated GAT-1 Performance Measures* [Report AFHRL-TR-73-72]. Air Force Human Resources Laboratory, Brooks Air Force Base.
- Rantanen, E. M., & Talleur, D. A. (2001). Measurement of pilot performance during instrument flight using flight data recorders. *International Journal of Aviation Research and Development*, 1(2), 89-102.
- Simple, C.A., Cotton, J.C. & Sullivan, D.J. (1981). *Aircrew Training Device*. [Report AFHRL-TR-80-58]. Air Force Human Resources Laboratory, Brooks Air Force Base.
- Talleur, D.A., Taylor, H.L., Emanuel, T.W. Jr., Bradshaw, G.L. & Rantanen, E.M., (2003), Personal computer aviation training devices: their effectiveness for maintaining instrument currency. *International Journal of Aviation Psychology*, 13(4), 387-399.
- Taylor, H. L., Talleur, D. A., Emanuel, T. W., Jr., Rantanen, E. M., Bradshaw, G. L., & Phillips, S. I. (2002). *Incremental training effectiveness of personal computer aviation training devices (PCATD) used for instrument training* (ARL-02-5/NASA-02-3). Savoy, IL: University of Illinois, Aviation Research Lab.
- Vreuls, D., Wooldridge, A.L., Obermayer, R.W., Johnson, R.M., Norman, D.A. & Goldstein, I. (1975). *Development and Evaluation of Trainee Performance Measures in an Automated Instrument Flight Maneuvers Trainer*. [Report NAVTRAEQUIPCEN 74-C-0063-1]. Human Factors Laboratory, Navel Training Equipment Center.