MULTIPLE HYPOTHESIS TESTING BY UNPAIRED SAMPLES FOR INDEXING CHANGEPOINTS IN A ROAD-INDUCED VIBRATION SIGNAL

László Róbert HÁRI[1]

Several methods have been developed for simulating non-stationary and non-Gaussian processes in packaging vibration testing, encompassing unique methods for the segmentation of road vehicle vibrations. However, only a limited number of those consider spectral characteristics. Thus, the current paper introduces a novel segmentation algorithm conducted in the time-frequency domain. The spectral characteristics obtained by short-time Fourier transform are compared by multiple hypothesis tests to find changepoints in a wheeled vehicle vibration sample. Different post hoc procedures are introduced against the inflating Type I. error.

Keywords: Multiple hypothesis testing, Segmentation, Spectrogram, post hoc procedures.

1. Introduction

Travelers [1], transported cargo and itself the vehicles [2] are subjected to road induced vibrations when travelling on rough pavements. A more systematic and theoretical analysis of possibly related topics is presented in [3], based on clustering of publications' keywords. Road vehicle vibrations (RVV) are often characterized by power spectral density (PSD) functions, which is a wide-spread procedure in packaging vibration testing (PVT) according to standards like ISO, ISTA, ASTM or MIL-STD, as discussed in [4–6]. PSD profiles are usually averaged from longer time-history records from different journeys. The averaging process moderates the effect of sporadic shocks in the spectrum [7]. In addition, the inverse Fourier transform with uniformly distributed random phase yields a Gaussian distributed random signal, which is stationary with respect to time [8–10]. However, stationary signals often contradict the real nature of RVV. It is shown that the non-Gaussian nature is caused by the non-stationarity of RVV [7]. Different approaches had been developed for non-stationary RVV simulations, which methods encompass changepoint detection. Changepoint detection has extensive literature [11], the frequently implemented approaches in PVT are introduced in the rest of the current section.

[1] Research assistant lecturer, Department of Logistics and Forwarding, University of Győr, Hungary, e-mail: hari.laszlo@sze.hu, https://orcid.org/0000-0001-5280-7744

The simplest detection methods utilize one or more moving statistics, such as moving mean, -RMS, -crest factor (CF), or -kurtosis (κ) [12]. A conjunction of the RMS drop-off distance and the CF is presented in [13]. The Bayesian detector [14] can find homogenous sections separated by changepoints in the International roughness index (IRI)- and Rutting measurement series. An at-most-one-change (AMOC) algorithm finds the changes in level, variance, autocorrelation between successive measurements. The Split spectra method [7] partially alleviates the stationarity of PSD-based simulations. Its initial form separates recordings into lower- and higher amplitude events, each simulated by an average PSD profile. Probability split spectra is a variant of the latter method. The PSD level is accounted for each frequency, and different spectrum quantiles can be found, such as they "*represent the probability that an encountered PSD level will be at or below the profile based on all data events recorded* - Ref. 4" (ibid.). Filtering is utilized in [15] to separate the rigid-body motion from structural high-frequency bursts for the case of railcar vibrations. Wavelet decomposition of road roughness records is applied in [16]. Wavelet-based Gaussian decomposition [4,17] uses continuous wavelet transform (CWT) to decompose an RVV into Gaussian components by an iterative process. A Cumulative sum - bootstrapping algorithm is responsible for the segmentation of the instantaneous magnitude of RVV in [18]. Railcar vibrations are analyzed in [19], revealing the usefulness of intrinsic mode functions (IMF) in describing frequency-type non-stationarities in random signals. Machine learning classifiers are developed in [20] for detecting shocks buried in RVV using different classification methods. Several predictors are used by the classifiers, such as: moving RMS, - CF, - κ, and DWT, HHT. Classifiers are assessed by Receiver operating characteristics and the specifically developed Pseudo-energy ratio/fall-out curve.

Different event-detection methods had been presented so far; still, only a few investigate spectral characteristics. The current article introduces a segmentation method designed to find similar regions of the STFT based on significance levels. The method has its strength in relying on only two but conventional thresholds, such as the significance limit and the time resolution.

2. Methods

Investigation of the autocorrelation function (ACF) determines a limit above which the signal is considered independent from previous periods. The STFT uses this limit. One vector is a discrete Fourier transform (DFT) with elements $a_{i,k}$ over the bandwidth $i = 0,1,...,100$ [Hz] at instants $k = 1,2,...,600$ [s]. The idea to highlight here is that $a_{i,k}$ has individual distributions at any k. Note that this is not a spectral density but a probability density of the DFT amplitudes, similarly to Fig. 3. The logarithm (base 10) of the STFT serves as input for the

MHT procedures. Two sample *t*-tests (MHT_t) and Wilcoxon rank sum tests (MHT_W) assess the similarities among adjacent sections of log-STFT. Afterward, the Bonferroni and Holm-Bonferroni adjustments are introduced against the inflating Type I. error. Hypotheses considered *truly* significant post hoc yield the borders of segments.

The RVV signal [21] is measured on a passenger car's cockpit sampled with 1 kHz during a 10 min long journey. Its autocorrelation is investigated in Fig. 1., and a one second limit is assumed sufficiently long to ensure a quasi-independent state of samples for the MHT. Thus, STFT in Fig. 2. is obtained with 1 s long windows. Because the distribution of the original STFT is heavily skewed towards lower amplitudes, a logarithmic transformation is applied to the STFT, yielding a more symmetric representation of the amplitude histograms per second in Fig. 3.

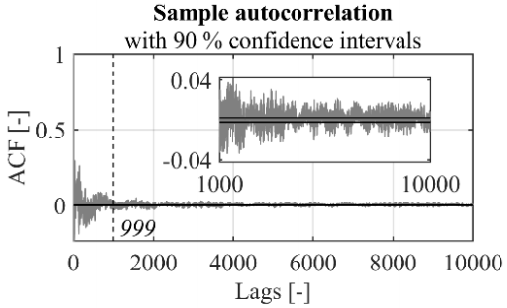

Fig. 1. Indication of the chosen limit of high autocorrelation at 999 lags; autocorrelation function of the signal (*gray*) and 90 % confidence interval (*black*).

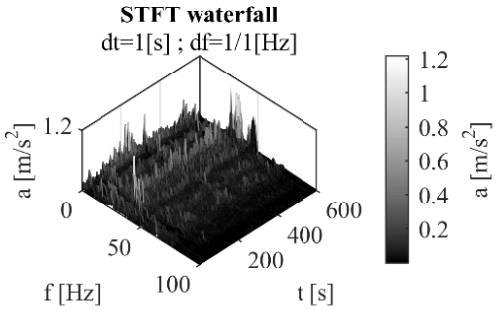

Fig. 2. Short-time Fourier transform of the vibration signal.

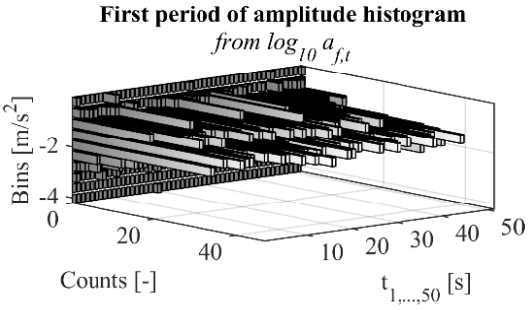

Fig. 3. Histogram of amplitudes (DFT elements) in the first 50 s of log-STFT vectors.

The STFT resulted in $K = 600$ sections, offering $J = 599$ comparisons per MHT. Despite the higher Nyquist frequency, the STFT are band-limited to the [0,100] Hz interval. Central tendencies are compared by the *t*- and rank sum tests to find similar segments among neighboring amplitude densities per second on the

preliminary significance level of α_0. The resulted series of p-values are compared to differently adjusted significance levels introduced in the α adjustment section.

2.1. Two-sample *t*-test

Statistical inference can be made about the null hypothesis of two samples having the same mean, using the two-sample *t*-test (also known as unpaired *t*-test). The alternative hypothesis formulates inequality among the means:

$$\begin{aligned} H_0 &: \bar{x} = \bar{y}; \\ H_A &: \bar{x} \neq \bar{y}. \end{aligned} \tag{1}$$

The two-sample *t*-test is a parametric test that compares location parameters of two independent samples. Assuming equal variances of populations, the test statistic under H_0 has Student's *t*-distribution with $\nu = n_x + n_y - 2$ degrees of freedom, and the pooled standard deviation replaces the sample standard deviations. Assuming unequal variances of the two samples, the test statistic under the H_0 has an approximate Student's *t*-distribution with the number of degrees of freedom given by Satterthwaite's approximation. This test is sometimes called Welch's *t'*-test [22].

2.2. Wilcoxon rank sum test

The Mann-Whitney U-test "*is the nonparametric equivalent of the t-test for means*" [23]. Albeit not the same procedure, the "*Wilcoxon rank sum test is equivalent to the Mann-Whitney U test*" [24]. This study is performed in MATLAB, which has dedicated command to the rank sum test. If *t*-test criteria cannot be entirely met, the nonparametric Wilcoxon rank sum test may be implemented to assess the null hypothesis that two samples belong to populations with equal medians.

2.3. Multiple hypothesis testing

This section investigates the assumptions of the *t*-test and formulates the MHT configuration. The Bartlett test of the null hypothesis assuming homoscedasticity returned a p-value of 0.00. That is, the log-STFT function does not have equal variances over time, which is not surprising. Still, two adjacent log-DFT vectors might have equal variances, which remains uninvestigated. This is done on purpose, as it is not suggested to automate the choice of test (parametric or nonparametric) based on the test of variances [25 p.298]. Also, there is not a consensus choosing a test in case of heteroscedasticity (ibid.). Since

already a log-transformation is introduced, the current paper stays at *t*-test, assuming equal variances. The Anderson-Darling tests of the log-DFT vectors showed normality 188 times; still, the *t*-test is robust to the assumption of normality [26]. Since the cause of outliers in the STFT is unknown and changes in spectral behavior is in scope, the paper proceeds with the two-sample *t*-test assuming equal variances. *Ceteris paribus* the Wilcoxon rank sum test is utilized in another MHT. The MHTs are formulated, as

$$\begin{aligned} H_0^{(j)} &: \bar{a}_{i,k} = \bar{a}_{i,k+1}; \\ H_A^{(j)} &: \bar{a}_{i,k} \neq \bar{a}_{i,k+1} \end{aligned} \tag{2}$$

for $k = 1,..., J$, expressing the test of central tendencies (subsequently *centers*) among the $k, k+1$-th vectors. Not rejecting $H_0^{(j)}$ shows two consecutive DFT vectors having the same centers, hence an association among the vectors. Conversely, rejecting $H_0^{(j)}$ in favor of $H_A^{(j)}$ indicates neighboring vectors not having the same centers, thus a dissimilarity among them. In the case of a significant result, a new segment is initiated. Current MHT are deployed on $\alpha_0 = 0.10$ preliminary significance limits. Although many tests are found significant preliminarily, not all of them may be considered *truly* significant due to α inflation resulting from simultaneous testing.

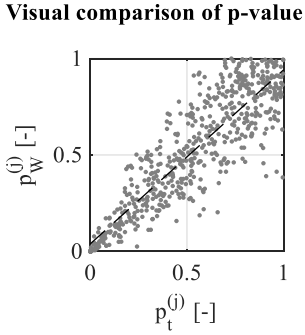

Fig. 4. Least squares fitted line (dashed) to emphasize similarities between p-values from *t*-tests $p_t^{(j)}$ and rank sum tests $p_W^{(j)}$.

2.4. α adjustment

Type of decisions in a single hypothesis test are summarized in Table 1. The probability of committing false statistical inferences increases when more

than one hypotheses are simultaneously tested, as is the case in MHT. Utilizing the same Type I. error rate in an increasing number of comparisons will increase the probability of at least one Type I. error.

Table 1

The decision framework

		Statistical inference	
		H_0 not rejected	H_0 rejected
Real fact	H_0 true	True negative ($1-\alpha$)	Type I. error (α)
	H_0 false	Type II. error (β)	True positive ($1-\beta$)

Probabilities in parentheses.

Adopting the same Type I. error level for m tests, the familywise error rate is

$$\alpha_{fw} = 1-(1-\alpha_0)^m, \tag{3}$$

also known as α inflation (Fig. 5.). Different α adjustment methods can be introduced as countermeasures to overcome the inflating likelihood of a Type I. error. The following sections discuss two of the possible methods, namely the Bonferroni and Holm-Bonferroni adjustments.

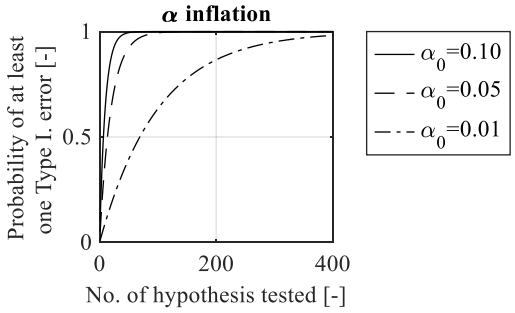

Fig. 5. Type I. error inflation

2.4.1. Bonferroni adjustment

The Bonferroni method [27] is a simple technique that makes it possible to make several comparison statements while ensuring that an overall confidence coefficient is preserved. The significance level is divided by the number of hypotheses tests, and each p-value is compared to the new significance level

$$\alpha_B = \alpha_0 / m. \tag{4}$$

The more hypotheses to be tested, the criterion gets more stringent and lowers the Type I. error per comparison, but also lowers the test's power.

2.4.2. Holm-Bonferroni adjustment

Holm adjustment [28] was subsequently proposed with less conservative character [29] and more power [30]. The method computes the significance levels α_{HB} depending on the rank of p-value. A step-down procedure is performed according to the ascendingly ordered $p^{(s)}$ value compared to successively increasing significance limits. The procedure similarly to [31] is as follows. The adjusted significance limit for the s-th hypothesis is

$$\alpha_{HB} = \frac{\alpha_0}{m - s + 1}, \tag{5}$$

and $H^{(1)},...,H^{(m)}$ are tested from the smallest to the largest p-values. The comparison stops at the first $p^{(s^*)} \geq \alpha_{HB}^{(s)}$ and $p^{(s^*)}$ with subsequent hypotheses are directly declared non-significant, viz. let s^* be the minimal index, such that

$$p^{(s^*)} \geq \frac{\alpha_0}{m - s^* + 1}, \tag{6}$$

all the hypotheses $H^{(1)},..., H^{(s^*-1)}$ are declared significant.

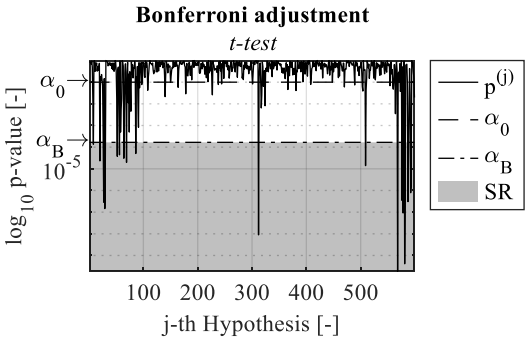

Fig. 6. Bonferroni adjustment α_B of the preliminary significance limit α_0 beneath $p_t^{(j)}$ yielding *SR* significance region.

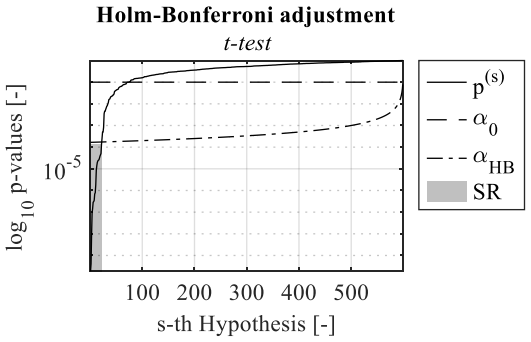

Fig. 7. Holm-Bonferroni adjustment α_{HB} of the preliminary significance limit α_0 beneath $p_t^{(s)}$ yielding *SR* significance region.

3. Results

The findings of current investigations are summarized as follows:

a) MHT_t and MHT_W behave similarly (Fig. 4.), yielding similar but not identical p-values for hypotheses at the same locations.
b) Given MHT_t, the Bonferroni and Holm-Bonferroni post hoc procedures yield the same hypotheses significant in $p_t^{(j)}$ and $p_t^{(s)}$. Similarly,
c) given MHT_W, $p_W^{(j)}$ and $p_W^{(s)}$ declares the same hypotheses significant.
d) Post hoc currently, one difference is noticeable between MHT_t and MHT_W, observable between Fig. 11. *b-c)* at $j = 8$, thus between t = (7,8] and (8,9] s.
e) Post hoc tests result in fewer segments compared to the preliminary tests.
f) Frequency modulation does not affect the presented method, e.g., 100-300 s.
g) Amplitude modulation affects the presented method, e.g., $j = 593$, thus between t = (592,593] and (593,594] s.

4. Discussion

This paper introduces variations on MHT procedures supplemented by post hoc tests to find segments in the spectrogram. Unfortunately, it is not allowed to make an *optimal* choice of tests in this paper because it would be a form of *p-hacking*. Importantly, the methods provide the following contributions.

a) The *t*-test assumes normality, the rank sum test requires symmetry in distributions. Despite the above-discussed assumptions assured to a certain extent, this is considered a good sign of the robustness of MHT_t and MHT_W, yielding preliminary only 15 differences.
b) MHT_t post hoc corrections do not yield different alternative hypotheses because the same p-values are in the different significance regions (Fig. 10.)
c) which is similarly valid for MHT_W per post hoc adjustments.
d) The scattering in Fig. 4. presents that the MHTs do not yield the same p-value series, as expected. This scattering is also observable in the low p-value regions between $p_t^{(8)}$ and $p_W^{(8)}$.
e) The moderated number of significant p-values post hoc is again an expected behavior since it is aimed to reduce the familywise error rate. A thought experiment in Fig. 10. shows the post hoc test's behavior as a function of number p-values for 599 and 5990 instances. With an increasing number of tests, post hoc procedures yield stricter limits.
f) It is not good observable from the data (Fig. 8.) but theoretically justifiable that the simple rearrangement of the elements in a log-DFT corresponding to, (e.g., at 34 s) would hold the same elements with different peak location. This should not yield a different amplitude distribution of the DFT vector.

g) Strong amplitude modulation occurring in the spectrogram can significantly manifest itself in the amplitude distribution of log-DFT elements (Fig. 9.), representing the algorithm's capability to detect amplitude modulation.

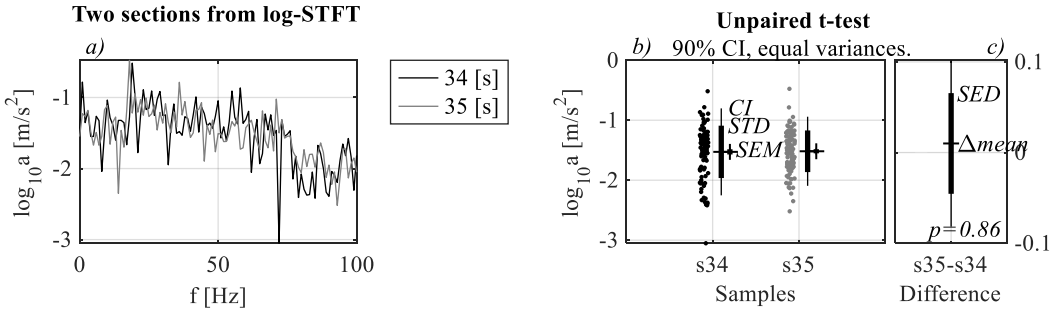

Fig. 8. The unpaired *t*-test between 34-35th log-DFT vectors showing a slight frequency modulation; STD (standard deviation), CI (confidence interval), SEM (standard error of the mean), SED (standard error of difference).

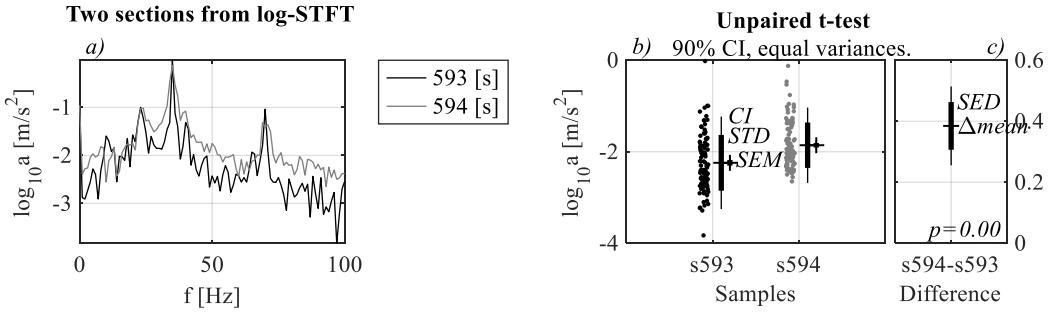

Fig. 9. The unpaired *t*-test between 593-594th log-DFT vectors showing amplitude modulation, notations similarly to Fig. 8.

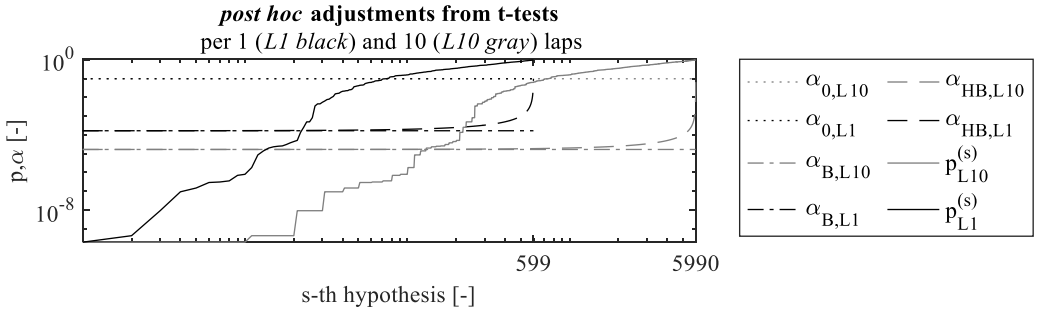

Fig. 10. A thought experiment of traveling the same rounds resulting in the same p-values per laps, but the significance limits from the Bonferroni and Holm-Bonferroni procedures differ at every round, e.g., $\alpha_{B,L1} \neq \alpha_{B,L10}$ or $p^{(s*)}_{HB,L1} \neq p^{(s*)}_{HB,L10}$.

Albeit the current findings apply only to the given 10 min long measurement (which cannot be referred to as a representative sample of RVV worldwide) the approach demonstrates promising results. Nevertheless, the method allows declaring segment-borders by well-established hypothesis tests instead of a heuristic *try-and-error* weighting of parameters.

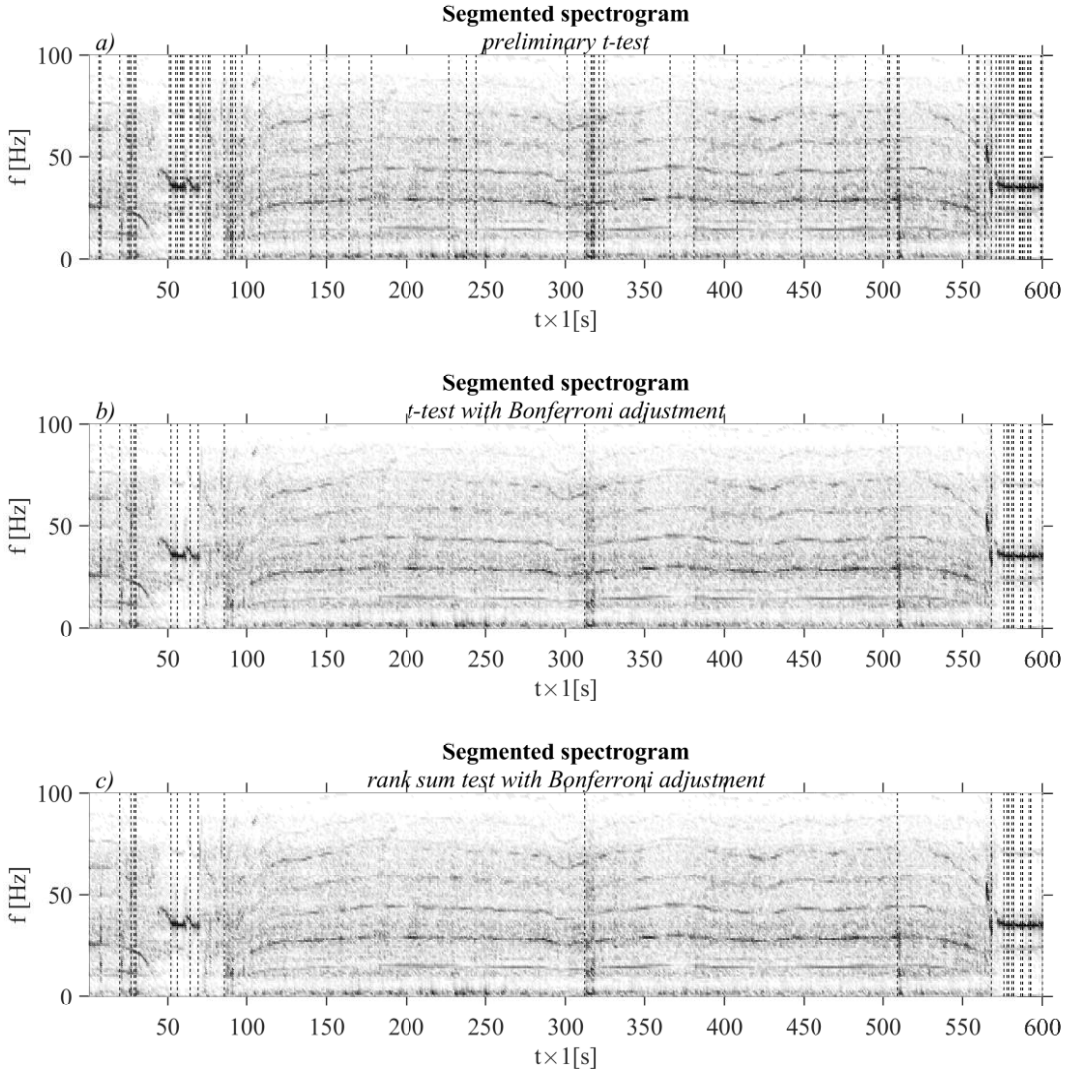

Fig. 11. Result of segmentation projected on the STFT surface. Dashed vertical lines denote segment borders; *a) t*-test on the preliminary significance limit α_0; *b) t*-test with Bonferroni adjustment; and *c)* rank sum test with Bonferroni adjustment.

Central tendencies have been assessed here, but future works aim at using hypothesis tests accounting for the shape of STFT vectors. Kolmogorov-Smirnov

type tests seem a possible choice on this purpose. Other error-controlling methods may also constitute future studies. The presented method assesses information among neighboring DFT vectors. An improved variant can include a guard against drift by consecutive amplitude or frequency modulation in the STFT.

6. Conclusions

Segments within a spectrogram can be found, in-between homogenous to a specific criterion. Different spectrums might simulate homogenous segments in PVT. Only a few segmentation methods investigate spectral properties, considering the discipline of PVT. Nevertheless, the number of techniques using objective hypothesis test is alarmingly low. Thus, the current paper introduced *Statistical Spectrogram Segmentation* (*3S*), an algorithm capable of detecting amplitude modulation in the time-frequency domain of RVVs. The segmentation is achieved by two MHT procedures supplemented by post hoc corrections. Overall, four implemented variations show good agreements. Therefore, the idea of MHT for RVV segmentation accounting amplitude modulation shows a straightforward and objective solution.

REFERENCES

[1] *Picu M.* The transmission of vibration through car seats. International Journal of Modern Manufacturing Technologies 2013;V:87–90.

[2] *Vasiu R-V, Melinte O, Vlădăreanu V, Dumitriu D.* On the response of the car from road disturbances. The Romanian Journal of Technical Sciences 2013;58:195–208.

[3] *Hári LR.* Keyword co-occurrence analysis of a packaging vibration testing relevant sample. 16th International Bata Conference for Ph.D. Students and Young Researchers, vol. 16, Zlín, Czech Republic: Tomas Bata University; 2020, p. 187–201. https://doi.org/10.7441/dokbat.2020.16.

[4] *Griffiths KR, Hicks BJ, Keogh PS, Shires D.* Wavelet analysis to decompose a vibration simulation signal to improve pre-distribution testing of packaging. Mechanical Systems and Signal Processing 2016;76–77:780–95. https://doi.org/10.1016/j.ymssp.2015.12.035.

[5] *Bonnin A-S, Nolot J-B, Huart V, Pellot J, Krajka N, Odof S, et al.* Decomposition of the acceleration levels distribution of a road transport into a sum of weighted Gaussians: application to vibration tests. Packag Technol Sci 2018;31:511–22. https://doi.org/10.1002/pts.2375.

[6] *Lepine J, Rouillard V, Sek M.* Review paper on road vehicle vibration simulation for packaging testing purposes: review of road vehicle vibration simulation for packaging testing. Packag Technol Sci 2015;28:672–82. https://doi.org/10.1002/pts.2129.

[7] *Kipp WI.* Random vibration testing of packaged-products: considerations for methodology improvement, Bangkok: 2008, p. 12.

[8] *Sek MA.* A modern technique of transportation simulation for package performance testing. Packag Technol Sci 1996;9:327–43. https://doi.org/10.1002/pts.2770090604.

[9] *Rouillard V, Sek MA.* Monitoring and simulating non-stationary vibrations for package optimization. Packaging Technology and Science 2000;13:149–56. https://doi.org/10.1002/1099-1522(200007)13:4<149::AID-PTS508>3.0.CO;2-A.

[10] *Rouillard V, Sek MA.* Synthesizing nonstationary, non-Gaussian random vibrations. Packaging Technology and Science 2010;23:423–39. https://doi.org/10.1002/pts.907.
[11] *Tartakovsky A, Nikiforov I, Basseville M.* Sequential analysis: Hypothesis testing and changepoint detection. Boca Raton, FL: Taylor & Francis Group, LLC; 2015. https://doi.org/10.1080/02664763.2015.1015813.
[12] *Bruscella B, Rouillard V, Sek M.* Analysis of road surface profiles. J Transp Eng 1999;125:55–9. https://doi.org/10.1061/(ASCE)0733-947X(1999)125:1(55).
[13] *Bruscella B.* Analysis and simulation of the spectral and statistical properties of road roughness for package performance testing. Master of engineering in mechanical engineering. Victoria University of Technology, 1997.
[14] *Thomas F.* Automated road segmentation using a bayesian algorithm. J Transp Eng 2005;131:591–8. https://doi.org/10.1061/(ASCE)0733-947X(2005)131:8(591).
[15] *Rouillard V, Richmond R.* A novel approach to analysing and simulating railcar shock and vibrations. Packag Technol Sci 2007;20:17–26. https://doi.org/10.1002/pts.739.
[16] *Wei L, Fwa TF, Zhe Z.* Wavelet analysis and interpretation of road roughness. J Transp Eng 2005;131:120–30. https://doi.org/10.1061/(ASCE)0733-947X(2005)131:2(120).
[17] *Griffiths KR.* An Improved Method for Simulation of Vehicle Vibration Using a Journey Database and Wavelet Analysis for the Pre-Distribution Testing of Packaging. University of Bath, 2012.
[18] *Rouillard V.* On the statistical distribution of stationary segment lengths of road vehicles vibrations. Proceedings of the World Congress on Engineering, vol. II, London: 2007.
[19] *Rouillard V, Sek MA.* The use of intrinsic mode functions to characterize shock and vibration in the distribution environment. Packag Technol Sci 2005;18:39–51. https://doi.org/10.1002/pts.677.
[20] *Lepine J, Rouillard V.* Evaluation of shock detection algorithm for road vehicle vibration analysis. Vibration 2018;1:220–38. https://doi.org/10.3390/vibration1020016.
[21] *Irvine T.* Vibrationdata.com. August 2004 Newsletter 2004:1–12. http://www.vibrationdata.com/Newsletters/August2004_NL.pdf (accessed November 24, 2020).
[22] *Welch BL.* The generalization of 'Student's' problem when several different population varlances are involved. Biometrika 1947;34:28–35. https://doi.org/10.1093/biomet/34.1-2.28.
[23] *Guthrie WF.* NIST/SEMATECH e-Handbook of Statistical Methods (NIST Handbook 151) 2020. https://doi.org/10.18434/M32189.
[24] Statistics toolbox: for use with Matlab. The MathWorks, Inc; 2005.
[25] *Motulsky H.* Intuitive biostatistics: a nonmathematical guide to statistical thinking. Fourth edition. New York: Oxford University Press; 2018.
[26] Moore DS. The basic practice of statistics. 5th ed. New York: W.H. Freeman and Co; 2010.
[27] *Bonferroni CE.* Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936.
[28] *Holm S.* A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 1979;6:65–70.
[29] *Lee S, Lee DK.* What is the proper way to apply the multiple comparison test? Korean J Anesthesiol 2018;71:353–60. https://doi.org/10.4097/kja.d.18.00242.
[30] *Kim H-Y.* Statistical notes for clinical researchers: *post-hoc* multiple comparisons. Restor Dent Endod 2015;40:172. https://doi.org/10.5395/rde.2015.40.2.172.
[31] *Chen S-Y, Feng Z, Yi X.* A general introduction to adjustment for multiple comparisons. J Thorac Dis 2017;9:1725–9. https://doi.org/10.21037/jtd.2017.05.34.