ecma INTERNATIONAL

Technical Report

ECMA TR/105

1st Edition / December 2012

A Shaped Noise File Representative of Speech

Rue du Rhône 114 CH-1204 Geneva T: +41 22 849 6000 F: +41 22 849 6001

is the registered trademark of Ecma International

COPYRIGHT PROTECTED DOCUMENT

Contents

Page

Introduction

Determination of headphone acoustic output for compliance with product safety regulations is described in EN 50332, which in turn references IEC 60268. Together, these documents describe three major characteristics of a recorded file that is to be used when measuring this output. These three characteristics are the spectrum, the crest factor, and the recording level. The spectrum is specified relative to pink noise, which has a flat spectrum when using constant percentage bandwidth filters, specifically third-octave filters out to 20 kHz.

Use of a shaped noise file is attractive because it can be described mathematically and has characteristics that are essentially the same considering any portion of the file beyond some reasonably short time scale. This means that long averaging times are not necessary, and a stable measurement can be made quickly. A purely mathematical description also means that the file can be generated by anyone, rather than relying on specific “golden” recordings.

Because EN 50332 is concerned with hearing safety, the crest factor is quite aggressive to encompass the behavior of certain types of music. However, in other cases, such as for power consumption testing, a noise file more representative of the typical behavior rather than the upper limit is desired. In addition, different content types, such as speech, are also of interest, for example listening to an audiobook or a podcast.

This Ecma Technical Report has been adopted by the General Assembly of December 2012.

A Shaped Noise File Representative of Speech

1 Scope

This Technical Report describes a digital shaped pink noise file representative of speech in two main characteristics, namely the spectrum and the crest factor. The spectrum is defined in third-octave band levels relative to pink noise up to the 8 kHz band, which provides a sufficient bandwidth for speech. The crest factor is defined at a 30 second time scale. The recording level of the file is not specified, and should be adjusted to match the amplitude, at the output of the headphone, of a typical audiobook or podcast when played on the device under test. This file is not meant to replace the existing file defined in EN 50332-1 for hearing safety.

2 References

For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

EN 50332-1, *Sound system equipment – Headphones and earphones associated with portable audio equipment – Maximum sound pressure level measurement methodology and limit considerations*

IEC 60268-1, *Sound system equipment*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

3.1
crest factor (CF)
Crest factor is the ratio of the largest absolute value in a time-varying signal to the root-mean-square (RMS) value of the signal. In this Technical Report, it will be expressed in decibels, and calculated by the following equation, where x is the time-varying amplitude.

$$\mathrm{CF} = 20 \cdot \log_{10}\left[\mathrm{Max}(|x|) \Big/ \mathrm{RMS}(x)\right] \quad (1)$$

4 Abbreviations

CF	crest factor
RMS	root-mean-square
SPL	sound pressure level

5 Spectrum

The spectrum is derived by taking the average spectrum of the TIMIT speech corpus files [2], which contain several hours of speech by 630 speakers of both genders, recorded at a sample rate of 16 kHz. A spline fit is then used to smooth the result. See Annex A for details on the derivation. The target spectrum, relative to pink noise, is listed in Table 1 by third-octave band, and is based on a 16 kHz sample rate. The tolerance on the spectrum is the same as that given in IEC 60268-1.

NOTE 1 Some analysis programs may not report a level for the 8 kHz band if the signal does not fill entire band, as is the case with a sample rate of 16 kHz.

NOTE 2 If the file is synthesized using a wider bandwidth, care must be taken when downsampling to 16 kHz, because the low-pass filter typically used to avoid aliasing will greatly alter the level of the resulting 8 kHz band. If a wider bandwidth is used during synthesis, the spectrum defined below may be extended at a slope of -2 dB per band. However, the final file must be limited to only the bands specified in Table 1 and a sample rate of 16 kHz.

Table 1 – Spectrum Relative to Pink Noise

Frequency (Hz)	Relative SPL (dB)
20	-49,8
25	-48,4
32	-47,3
40	-46,2
50	-44,6
63	-40,8
80	-34,0
100	-26,4
125	-20,7
160	-17,2
200	-14,1
250	-11,0
315	-7,7
400	-4,0
500	-1,1
630	0,0
800	-1,1
1 000	-3,4
1 250	-5,7
1 600	-7,8
2 000	-9,9
2 500	-11,5
3 150	-12,4
4 000	-13,1
5 000	-14,3
6 300	-16,0
8 000	-18,1

6 Crest Factor

The crest factor, calculated over successive, non-overlapping windows of 30 seconds in duration, shall be 24 dB ± 1 dB in each complete window. See Annex B for details on the derivation.

NOTE 1 An power function is one possible implementation to alter the crest factor of the shaped noise file once the band levels had been adjusted to the target. This involves a non-linear gain applied to the amplitudes of each sample, so that those samples that are farther from zero are increased in magnitude more than those closer to zero, thus increasing the crest factor.

NOTE 2 The crest factor adjustment can interact with the band levels, requiring a precompensation of levels or an iterative process to meet both requirements simultaneously.

Annex A
(informative)

Basis of Target Spectrum

Because the TIMIT files are long and contain multiple speakers, the spectra of the files are expected to be similar, and this is shown to be the case in Figure A.1, where an arbitrary offset on the amplitude scale is used. To produce a representative spectrum, the spectra of all TIMIT files were averaged, and a spline fit used to slightly smooth the result, as shown in Figure A.2. The spline fit was then adjusted to have a maximum value of 0 dB to produce the values in Table 1 above.

A comparison of the spectra of the IEC 60268-1 file and the speech-representative file is shown in Figure A.3. Note that speech is much more concentrated into the mid-frequency range than the IEC file. Also note that the spectral content of the speech file stops at the 8 kHz band, while that of the IEC file continues to 20 kHz; this reflects that the IEC file is intended to be representative of music, which requires a wider bandwidth for reproduction.

Figure A.4 shows a spectrum comparison between one of the TIMIT files and the synthesized shaped noise file. The stepwise nature of the synthesized file spectrum can be seen, as the level of each band is adjusted individually to meet the desired target.

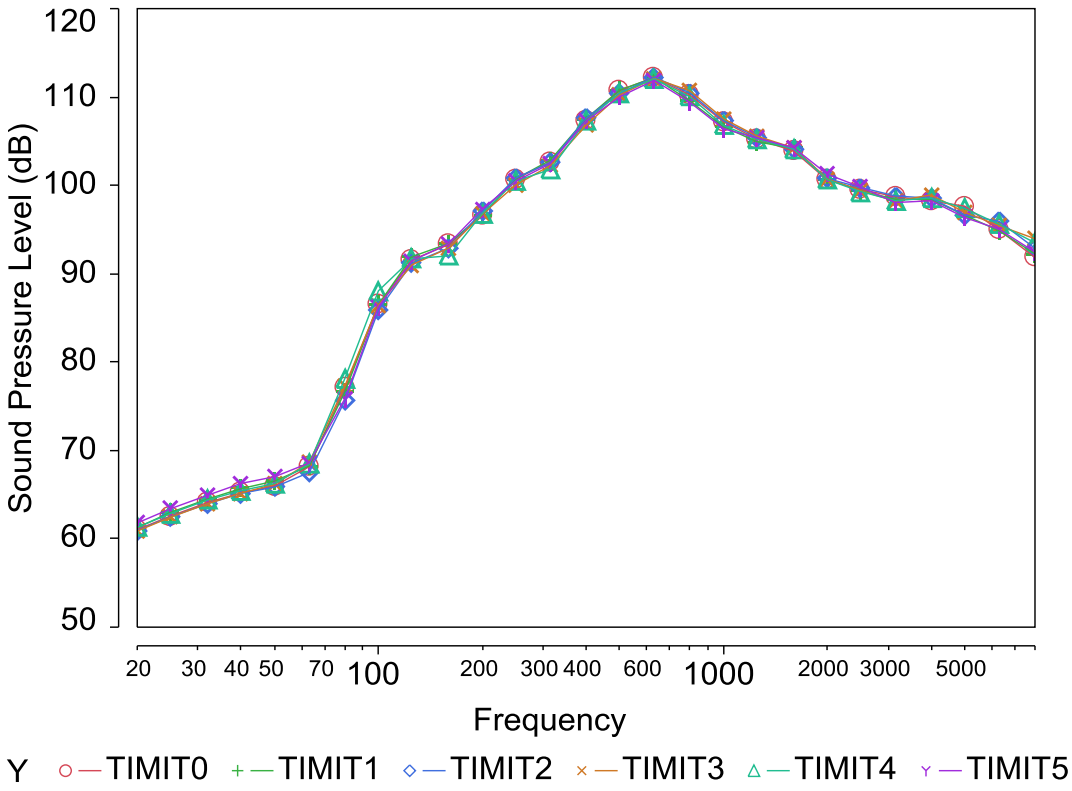

Figure A.1 – Third-octave spectra of individual TIMIT files (arbitrary offset).

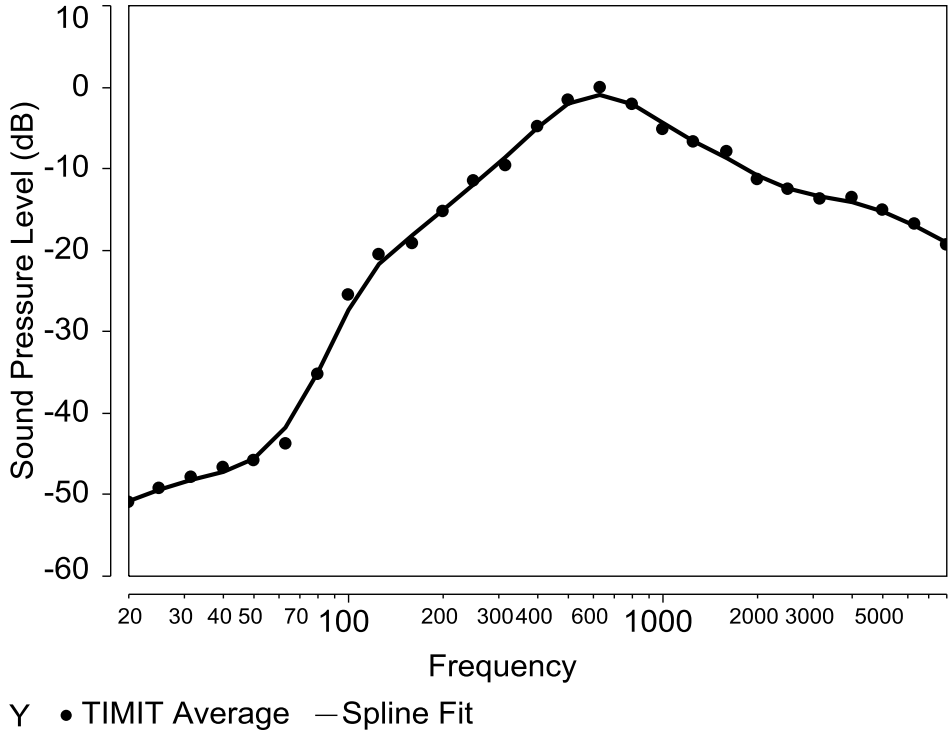

Figure A.2 – Average third-octave TIMIT spectrum and spline fit (arbitrary offset).

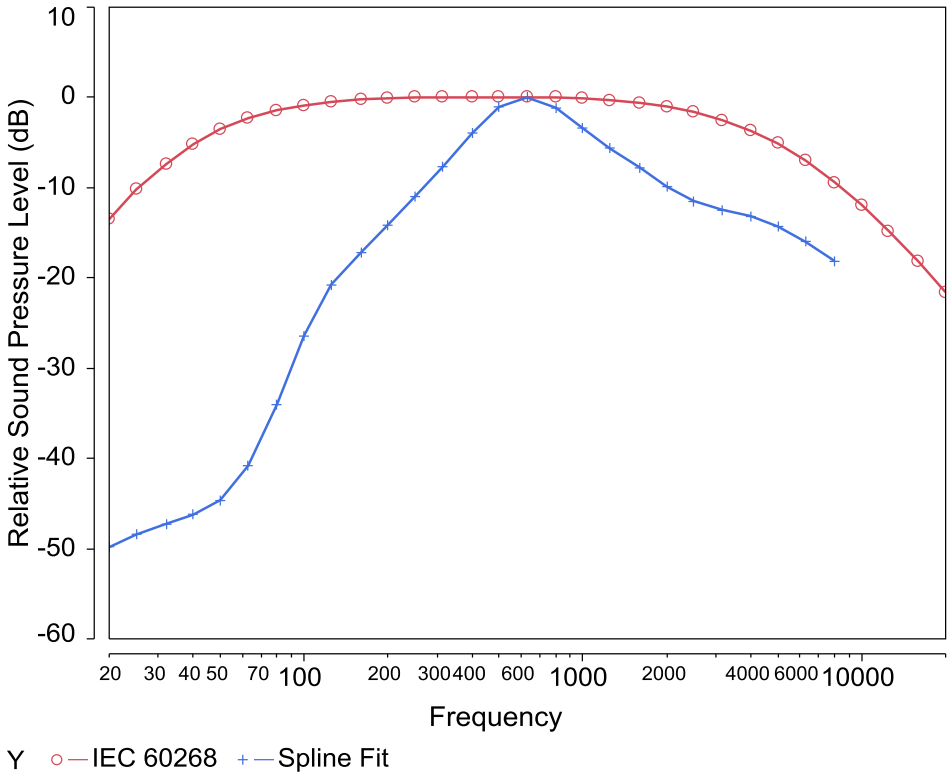

Figure A.3 – Comparison of IEC 60268 and speech third-octave spectra, relative to pink noise.

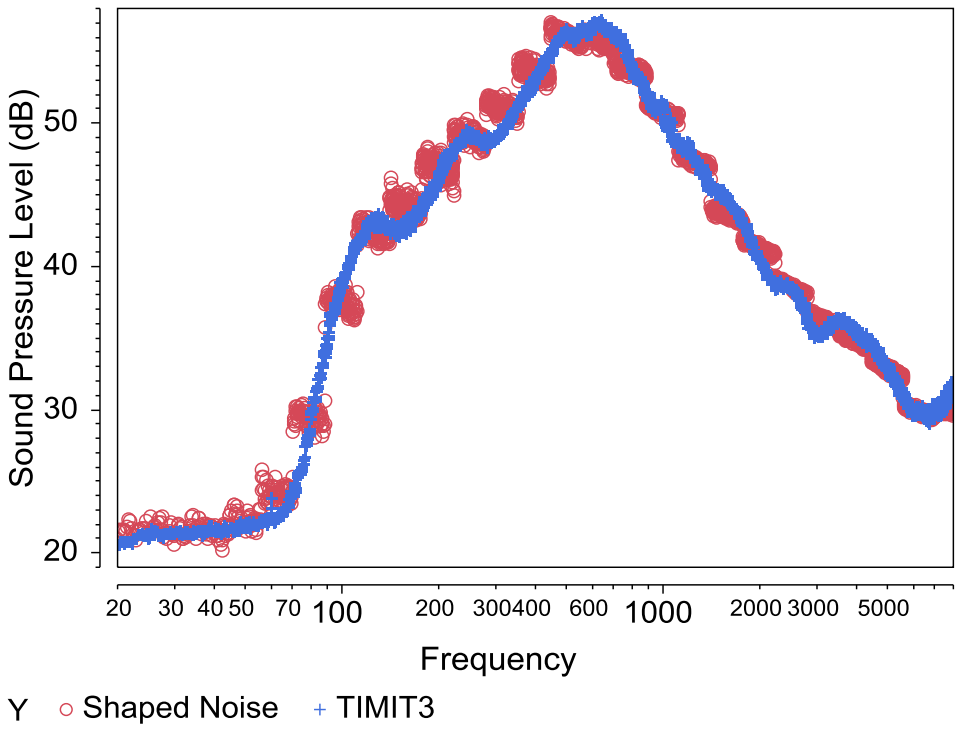

Figure A.4 – FFT comparison of the speech-representative shaped noise file and the TIMIT3 file (gains adjusted so traces overlap).

Annex B
(informative)

Basis of Target Crest Factor

Because crest factor is dependent on the single point of maximum amplitude, any segment of a recording which does not include this point cannot have the same crest factor as that obtained from analysis of the recording as a whole, except by coincidence. This also means that a single outlier value will determine the crest factor of the entire file, regardless of how the rest of the file behaves. To examine this effect, the crest factor of the TIMIT files was analyzed using a variety of time windows. Figure B.1 shows the crest factor for the first TIMIT file, which is about one hour in length, as a function of the analysis window length. In this plot the green line indicates the mean value while the red box indicates the center two quartiles of data. It can be seen that the mean crest factor increases with increasing window length.

To understand this behavior, consider Figure B.2 and Figure B.3, which are the RMS and maximum absolute value, respectively, in the same analysis windows. In Figure B.2, the mean RMS at any window length is remarkably stable. At short window lengths, many RMS values are averaged together across the entire file, while at long window lengths many points are averaged to create each RMS. The behavior is quite different for the maximum absolute value. In Figure B.3, the overall maximum absolute value in the entire file is the point at the top of each column (with a value over 16,000), but at short window lengths there are many other local maxima found which are lower than the overall maximum. As the window length increases, so does the chance of encountering a relatively high local maximum value, and thus the average of the maximum absolute value increases with window length. This is what leads to the increase in average crest factor with window length in Figure B.1.

This behavior is summarized in Figure B.4 for the TIMIT0 file, and the results are similar for the other TIMIT files. The curve fit for Figure B.4 is given below, where window length is in seconds.

$$\text{Mean CF} = 18{,}24 + 3{,}70 \cdot \log_{10}(\text{window length}) \qquad \text{(A.1)}$$

This means that it is impossible to define a crest factor which unambiguously characterizes speech, since increasingly more extreme outliers tend to be found with longer window lengths. In order to define a representative value for the crest factor, an analysis was done of the statistical differences in the mean crest factor (at a 95% confidence level) among the six TIMIT files as a function of window length. At a window length of one second, three statistically significant groups are found within a range of means of 0,32 dB. This small difference reaches statistical significance due to the large number of points in the sample populations. At a window length of 10 seconds, only two statistically significant groups are found, with a range of means of 0,45 dB. At a window length of 30 seconds, the mean crest factor of the TIMIT files are statistically indistinguishable across a range of means of 0,43 dB. Thus, 30 seconds is long enough to render statistically negligible any differences in the various TIMIT files. However, it is still short enough to obtain many averages from a wide variety of speech recordings, yet long enough that exposure to a sound for this amount of time would give a listener a good idea of the nature of the sound. In addition, any glitch or pop in the recording can only affect the crest factor in one 30 second segment, which would mitigate the impact on calculation of the average crest factor in a recording lasting several minutes or more. A 30 second analysis window is thus suggested as the representative value, resulting in a crest factor of 24 dB (rounded to the nearest decibel) for every TIMIT file.

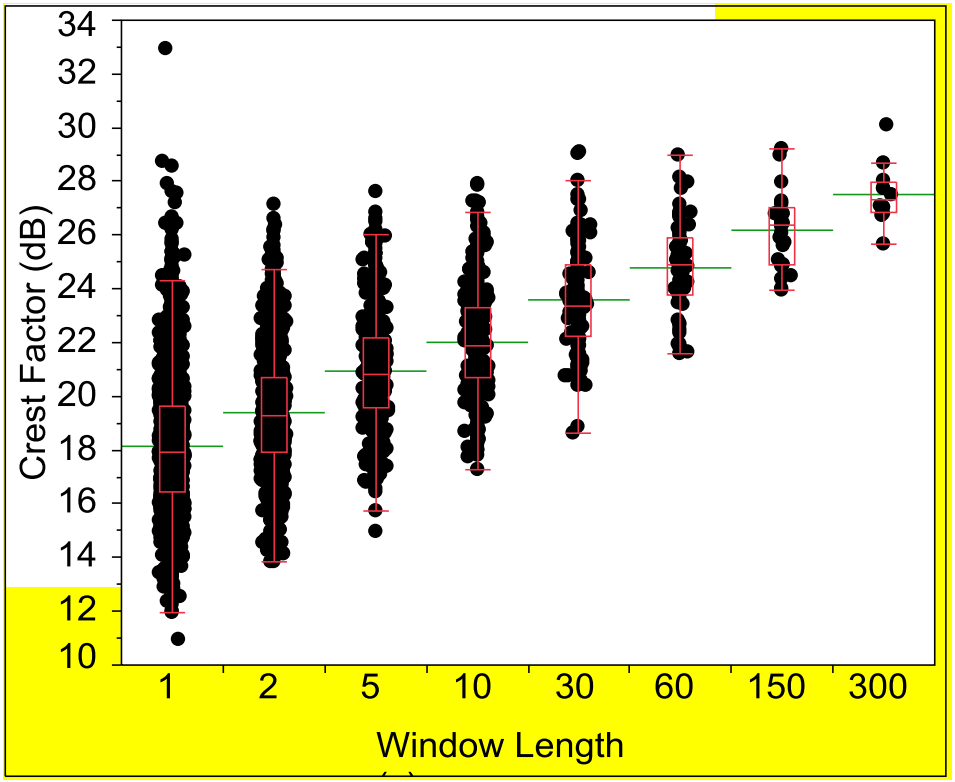

Figure B.1 – Crest factor of TIMIT0 as a function of analysis window length. Green line indicates the mean value.

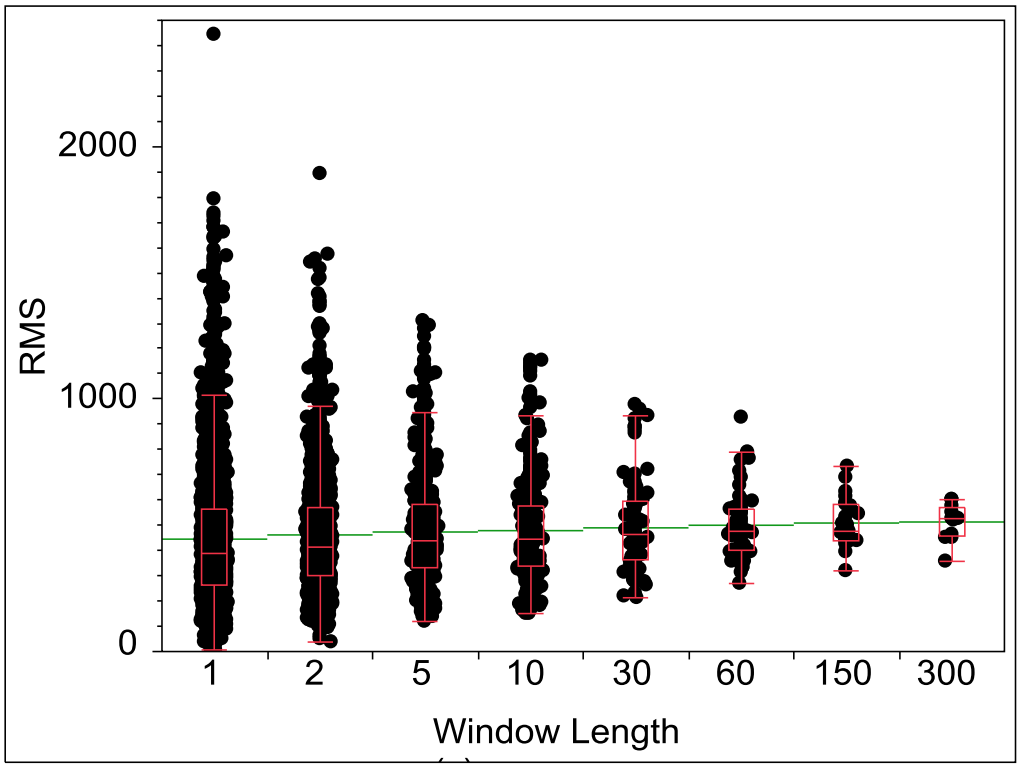

Figure B.2 – RMS of TIMIT0 as a function of the analysis window length. The units of the RMS are in 16 bit levels.

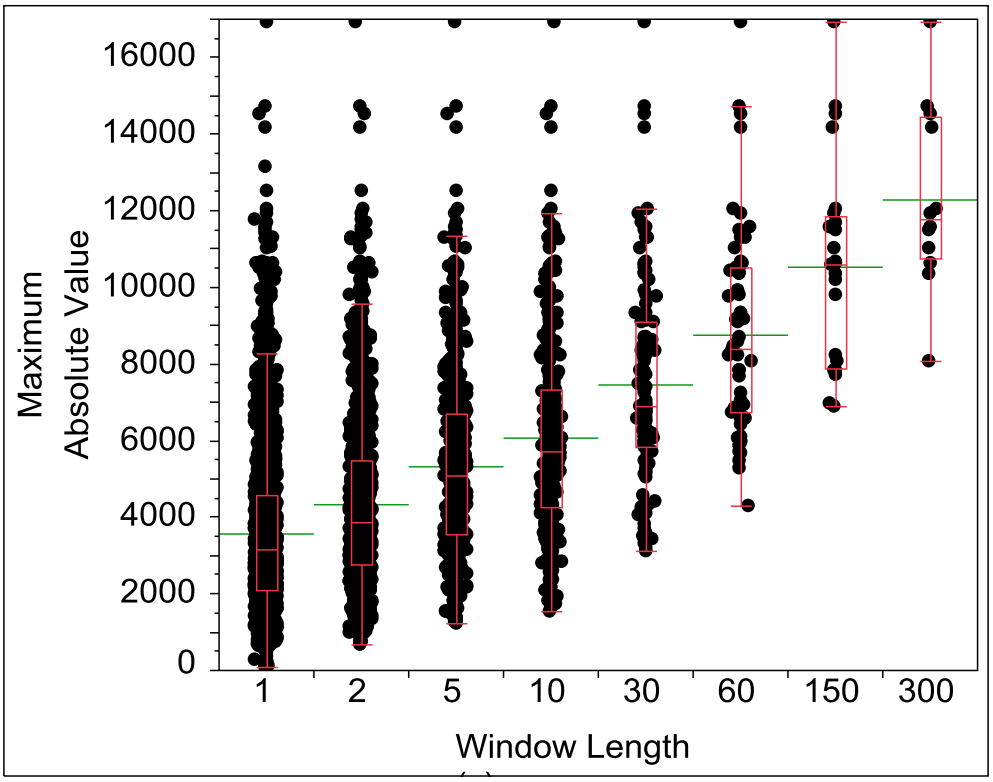

Figure B.3 – Maximum absolute value of TIMIT0 as a function of analysis window length. The units of the absolute value are in 16 bit levels.

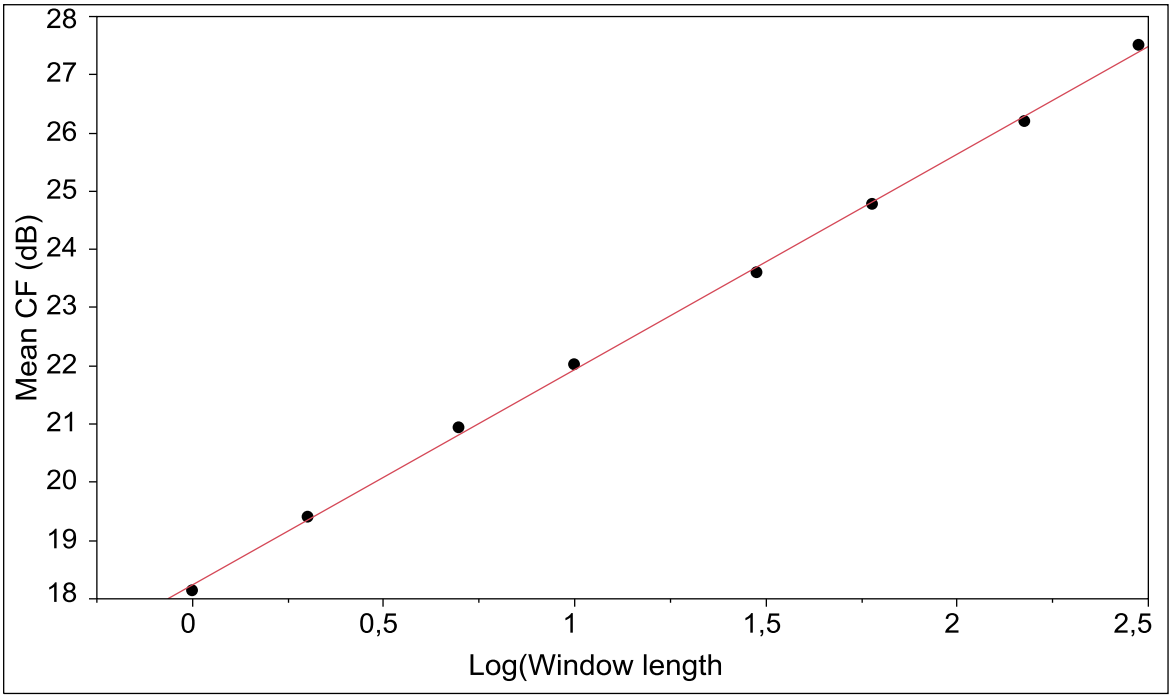

Figure B.4 – Mean crest factor as a function of the log of the analysis window length for TIMIT0.

Bibliography

[1] Baugh, Eric, “A shaped noise file representative of speech”, proceedings of Inter-Noise 2012

[2] Garofolo, John S., *et al.*, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, 1993