On the assessment of subjective response to tonal content of contemporary aircraft noise

Torija Martinez, AJ, Roberts, S, Woodward, R, Flindell, IH, McKenzie, A and Self, RH

http://dx.doi.org/10.1016/j.apacoust.2018.11.015

ON THE ASSESSMENT OF SUBJECTIVE RESPONSE TO TONAL CONTENT OF CONTEMPORARY AIRCRAFT NOISE

Antonio J. Torija[a]*, Seth Roberts[b], Robin Woodward[b], Ian. H. Flindell[b], Andrew R. McKenzie[b] and Rod H. Self[a]

[a]ISVR, University of Southampton, Highfield Campus, Southampton, SO17 1BJ UK

[b]Hayes McKenzie Partnership Ltd, Unit 3 Oakridge Office Park, Salisbury, SP5 3HT UK

Author to whom correspondence should be addressed. Electronic mail: A.J.Martinez@soton.ac.uk

Tel.: +44 (0)23 8059 2276

Abstract

The Effective Perceived Noise Level (EPNL) is the primary metric used for assessing subjective response to aircraft noise. The EPNL comprises calculation of the Perceived Noise Level (in PNdB), and takes into account flyover duration and the presence of pure tones to arrive at an adjusted EPNL value. With the presence of a single significant tone, EPNL has been found to be reasonably effective for the assessment of aircraft noise annoyance. Several authors have, however, suggested that EPNL is not capable of quantifying the subjective response to aircraft noise that contains multiple complex tones. The noise source referred to as "Buzz-saw" noise is a typical example of complex tonal content in aircraft noise with an important effect on both cabin and community noise impact. This paper presents the results of a series of listening tests where a number of participants were exposed to samples of aircraft noise with six variants of aircraft engines, assumed representative of the contemporary twin engine aircraft fleet. On the basis of the findings of these listening tests, the Aures tonality method significantly outperforms the EPNL tone correction method when assessing the subjective response to aircraft noise during take-off with the presence of multiple complex tones. The participants reported 'high pitch' as one of the least preferable aircraft noise characteristics, and consequently, the psychoacoustics metric Sharpness was found to be another important contributor to subjective response to the noise of two specific aircraft engine groups (out of the six considered). The limitations of Aures tonality are discussed, in particular for aircraft noise with both a series of complex tones spaced evenly across the frequency spectrum with relatively even sound levels and less subjectively dominant single frequency tones (compared to broadband noise). In line with these limitations, further work is proposed for more effective assessment of subjective response to aircraft noise containing significant tonal content in the form of numerous closely spaced or other complex tones.

Keywords: Aircraft noise; Complex tones; Tonality methods; Sharpness; Subjective response; Listening tests.

1. Introduction

Over the past few decades air traffic has experienced a significant growth, and this trend is projected to continue in the long term, driven by global GDP growth [1]. As a consequence, millions of people around the world already exposed to aircraft noise are likely to be subject to an increased aircraft noise exposure, particularly near airport infrastructures. In fact, it is recognised that noise is very often the most limiting factor in respect of airport development and mitigation strategies are often required to minimise noise impacts from airports [2]. When large residential areas are affected by aircraft noise it is not always easy to quantify the best noise mitigation strategy and research has been carried out to assist in determining the effectiveness of noise mitigation strategies [3-4]. Despite the many efforts made by the different stakeholders to limit or reduce aircraft noise impacts, noise pollution around airports continues to be a major health problem, with health effects recognised such as cardiovascular disease [5], sleep disorders with awakenings [6], and hypertension ischemic heart disease [7-8].

At a fleet level, the aircraft noise impact is assessed using noise contours of different exposure metrics (e.g. DNL, DENL, $L_{eq,16h}$, etc), under the assumption of the exposure-response relationship between the noise exposure and the percentage of (highly) annoyed persons [9]. At a vehicle level, the Effective Perceived Noise Level (EPNL), developed and introduced in the 1960s [10-13], is the metric currently used for aircraft noise certification purposes [14]. EPNL is calculated according to a procedure described by the Federal Aviation Administration (FAA) [15], which accounts for the Perceived Noise Level [16] and duration effects, and also applies a tonal penalty based on the level of the strongest protruding tone [17].

Although there is no consensus about the effects of (aircraft) tonal noise on human response, there is clear evidence that the tonal content is a major contributor toward the

perceived annoyance due to aircraft noise [18-19]. In this sense, Berckmans et al. [20] found that the (perceived) quality of the sound from an aircraft is mainly determined by the presence or absence of tonal components, with the loudness and frequency of the tonal components as key factors influencing noise annoyance. White et al. [21] observed a significant reduction in noise annoyance when the major tonal components were removed from aircraft noise samples. Currently, aircraft tonal noise annoyance is assessed by the tone correction method included in the derivation of EPNL. However, several authors [18,22-24] have claimed that there is a need for a metric that better captures more complex aircraft noise spectral characteristics.

As described above, the EPNL Tone Correction was developed to account for possible isolated tonal prominence. For that reason, this paper is firstly aimed at assessing whether the EPNL Tone Correction appropriately assesses the human response to tonal content in contemporary twin engine aircraft noise, which may well differ from that of aircraft in use at the time of its development. Secondly, since the EPNL Tone Correction was devised, more advanced methods of tonal analysis have been developed [25-29]. Most of these methods use much more refined frequency resolution analysis, overcoming the deficiencies of the third-octave banding for tonal identification [30] used in the calculation of the EPNL tone correction factor. Aures Tonality [26] is not limited to the maximum tonal emergence, and takes into account the presence of multiple tones. This provides a reasonably high correlation between the annoyance due to aircraft tonal noise and Aures Tonality [18,19]. Therefore, this paper also investigates whether Aures Tonality can improve the assessment of human response in relation to tonal content of contemporary aircraft. Thirdly, the noise signature of some future aircraft designs and engine developments, such as Ultra-High Bypass Ratio (BPR) turbofans [31], Counter Rotating Open Rotor (CROR) engines [23,30,32], or Distributed Electric Propulsion systems [33], will be likely to have a significant complex tonal content. For this reason, this paper discusses the limitations of current tonality assessment methods, and

suggests potential improvements for appropriately assessing the subjective response to aircraft noise with significant complex tonal content.

The research objectives stated above were addressed through the development of a series of listening tests where a number of participants were exposed to a series of samples of aircraft noise. These aircraft noise samples, recorded during take-off operations, contain a representative sample of contemporary twin engine aircraft of different size and entry-into-service (EIS) date, with six distinct aircraft engine variants. The engines selected are representative of smaller aircraft: Airbus A320 family (CEO and NEO) and Boeing 737 (800 and MAX); and larger aircraft: Boeing 757, 767, 777 and 787. From a perceptual point of view, three individual components are identified in the aircraft noise recordings, broadband noise (mainly jet and airframe noise), isolated tonal components (i.e. blade-passing frequency tone – BPF – generated by turbofans), and complex tones, including a series of numerous harmonic partials with a low fundamental frequency (commonly known as buzz-saw noise, BSN) [34].

Two points should be noted in relation to the samples that have been used in the listening tests. Firstly, samples were taken from recordings during take-off only since this is the time when engines are likely to be at the highest thrust setting over the duration of a flight and will therefore produce the most noise. Take-off noise may not be entirely representative of all aircraft noise but the aim was to capture recordings representative of the acoustic signature produced by the engine. Secondly, recordings were made at a single airport and two thirds of the engine variants selected for the listening test were from larger aircraft which may seem to over-represent these aircraft. However, the samples were selected to provide a varied mix of acoustic signatures representing a range of commonly used engine variants covering a broad timeline of engine development. The actual numbers of different aircraft represented by

the samples used in the listening tests are not intended to be representative of the mixture (percentage) of aircraft in use within modern airport fleets.

This paper is structured as follows: Section 2 presents a brief overview of the EPNL Tone Correction and the Aures Tonality method; Section 3 describes the equipment, stimuli and methodology used for the development of listening tests; Finally, in Section 4 and 5 the experimental results are presented and discussed respectively.

2. Metrics used for assessing human response to aircraft tonal noise samples

Although there are several methods for the detection, quantification and subjective response to tonal components [25-29], as discussed above, this paper focuses on two tonality metrics: (i) the EPNL Tone Correction, currently used for assessing spectral irregularities in aircraft noise [1]; and (ii) Aures Tonality, which has been found to be able to appropriately assess the presence of both isolated tones and harmonic series covering a wide band of frequencies [26].

Sharpness is a well-known metric used to measure the high frequency content of a sound [35]. The high frequency components have been found to play a significant part in the preference rating of aircraft noise [36]. Moreover, in an experiment assessing annoyance when exposed to aircraft noise [20], the participants reported the tonal components in the high frequency region (above 4000 Hz) as extremely annoying.

2.1. Effective Perceived Noise Level – Tone correction factor

The calculation of the EPNL Tone Correction requires the analysis of the sound-levels in third-octave frequency bands from 80 Hz to 10 kHz (L), in 0.5 s time steps. The tone correction is based on the position of the tone within the frequency spectrum and its excess sound-level over the sound-levels of the adjacent third-octave frequency bands (as explained below). The first criterion for the detection of tonal components is identifying those third-octave bands with sound-levels at least 5 dB higher than either of their 2 adjacent bands. After an iterative process for smoothing the original third-octave frequency spectrum (L), by averaging the sound-levels in adjacent bands [15 (steps 3-7)], a so-called background third-octave frequency spectrum (L') is calculated. The tonal prominence (or Level Difference, F) in each third-octave band is calculated as

$$F = L - L' \tag{1}$$

Then, F[1] is transformed into a tonal correction factor (C) depending on the third-octave frequency band. As shown in Fig. 1, for third-octave bands below 500 Hz or above 5 kHz, C ranges between 0 and 3.33 dB; while for third-octave bands between 500 Hz and 5 kHz, C ranges between 0 and 6.67 dB. The tone correction factor for each 0.5 s time step is the highest value of C throughout the third-octave band spectrum [15]. In this paper the overall EPNL tone correction factor was calculated as the average value of all the 0.5 s time steps.

[1] Note that only values of F equal or greater than 1.5 dB are considered further.

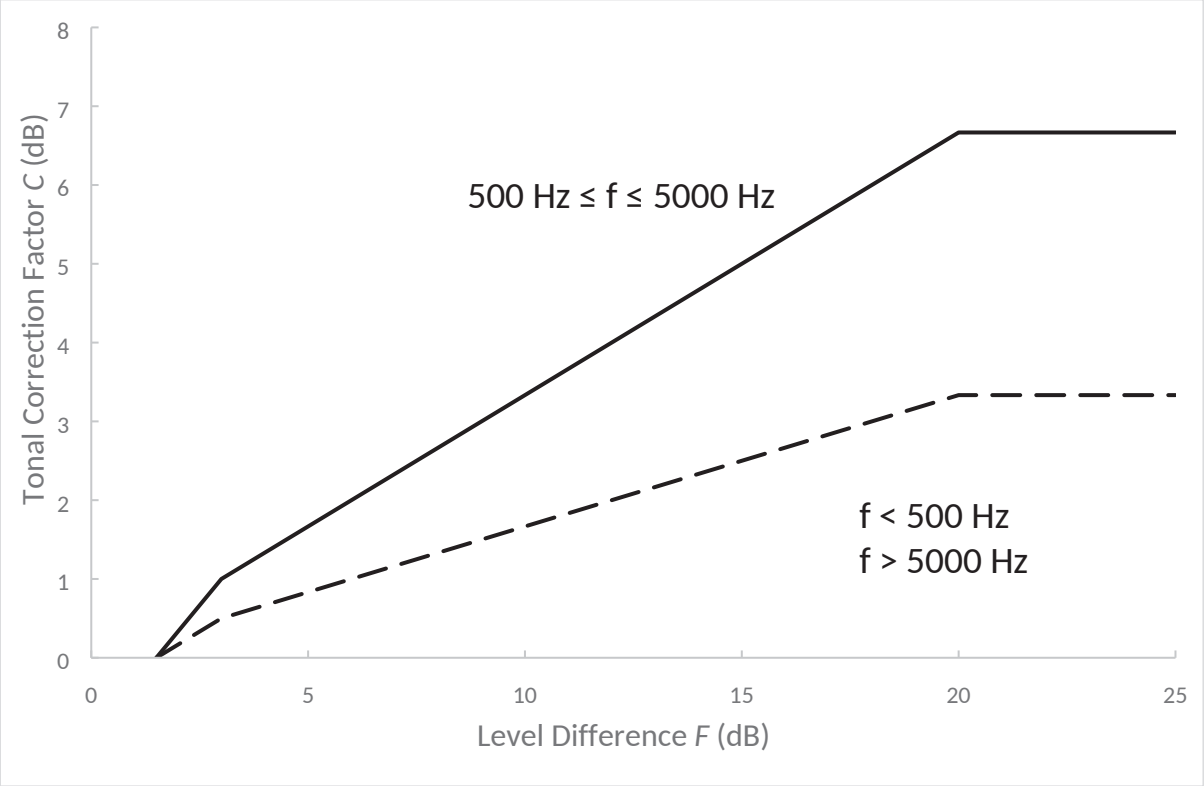


Figure 1. Tonal Correction Factor *C* as a factor of the Level Difference *F*.

2.2. Aures Tonality

Aures Tonality [26] uses an 80 ms signal window narrowband analysis to identify tones, and then applies weighting functions based on the bandwidth (Eq. 2), centre frequency (Eq. 3) and prominence (Eq. 4) of each tonal component.

$$w_1(\Delta z_i) = \frac{0.13}{\Delta z_i + 0.13} \tag{2}$$

$$w_2(f_i) = \left(\frac{1}{\sqrt{1 + 0.2\left(\left({f_i}/{700}\right) + \left({700}/{f_i}\right)\right)^2}} \right)^{0.29} \tag{3}$$

$$w_3(\Delta L_i) = \left(1 - e^{\left(\frac{-\Delta L_i}{15}\right)}\right)^{0.29} \tag{4}$$

where Δz_i is the bandwidth of the detected tonal component i expressed in Bark (i.e. as a fraction of the critical bandwidth), f_i is the frequency of the tonal component i in Hz, and ΔL_i is the level of the tonal component i above the broadband masking noise, as proposed by Terhardt et al. [25 (Eqs. 4-8)]. Figs. 2-4 show the change in weighting functions $w_1(\Delta z_i)$, $w_2(f_i)$ and $w_3(\Delta L_i)$ as a function of Δz_i (Eq. 2), f_i (Eq. 3) and ΔL_i (Eq. 4) respectively.

Then, an overall weighting function is derived by combining the weighting functions described in Eqs. 2-4:

$$w_T = \sqrt{\sum_{i=1}^{n}\left[\left(w_1(\Delta z_i)^{\frac{1}{0.29}}\right)\left(w_2(f_i)^{\frac{1}{0.29}}\right)\left(w_3(\Delta L_i)^{\frac{1}{0.29}}\right)\right]^2} \quad (5)$$

Aures tonality (K) in tu (tonality unit) is calculated as

$$K = c \cdot w_T^{0.29} \cdot w_{Gr}^{0.79} \quad (6)$$

where w_{GR} is a weighting function that accounts for the overall loudness of tone to noise ratio, calculated as

$$w_{GR} = 1 - \frac{N_{GR}}{N} \quad (7)$$

with N_{GR} as the loudness of the broadband noise component, and N as the total loudness of the sound,

and where $c = 1.09$ is a calibration constant, so that a 1 kHz pure tone with a sound-level of 60 dB has a tonality of 1 tu.

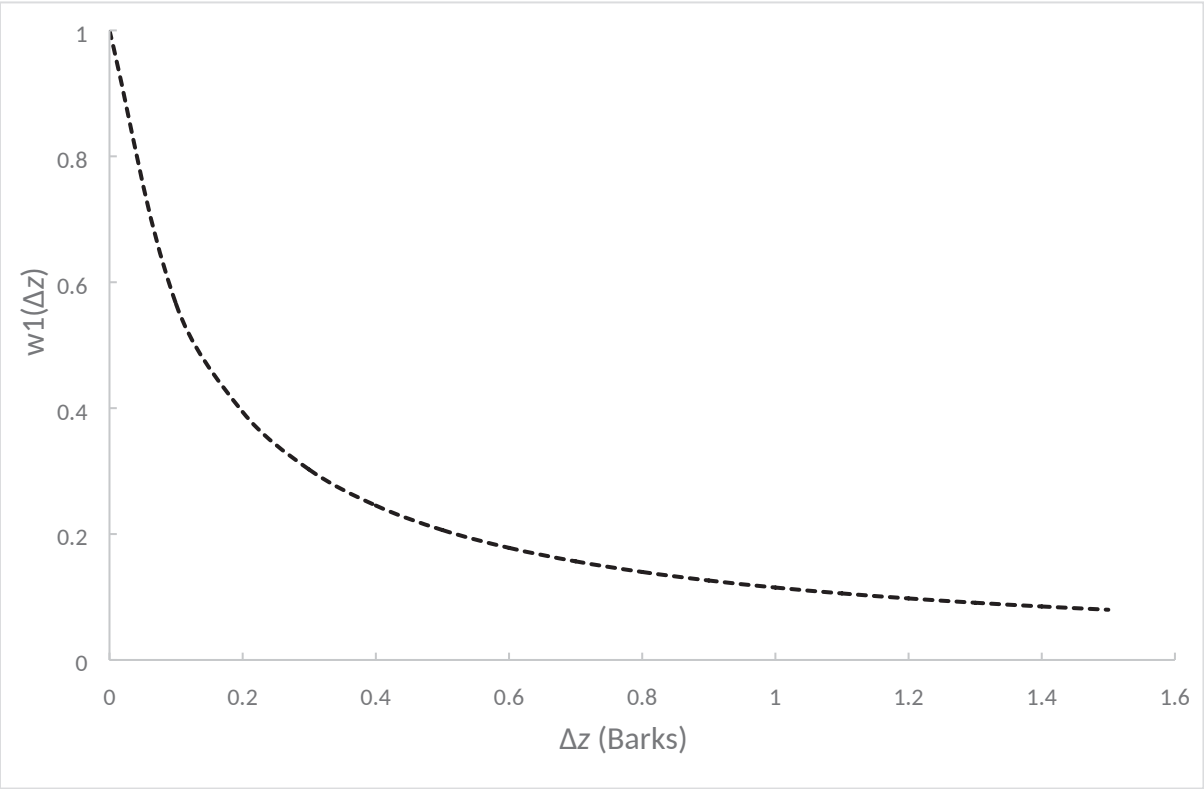


Figure 2. Weighting function $w_1(\Delta z)$ as a function of the bandwidth (Δz) in Barks of the detected tonal component (Eq. 2).

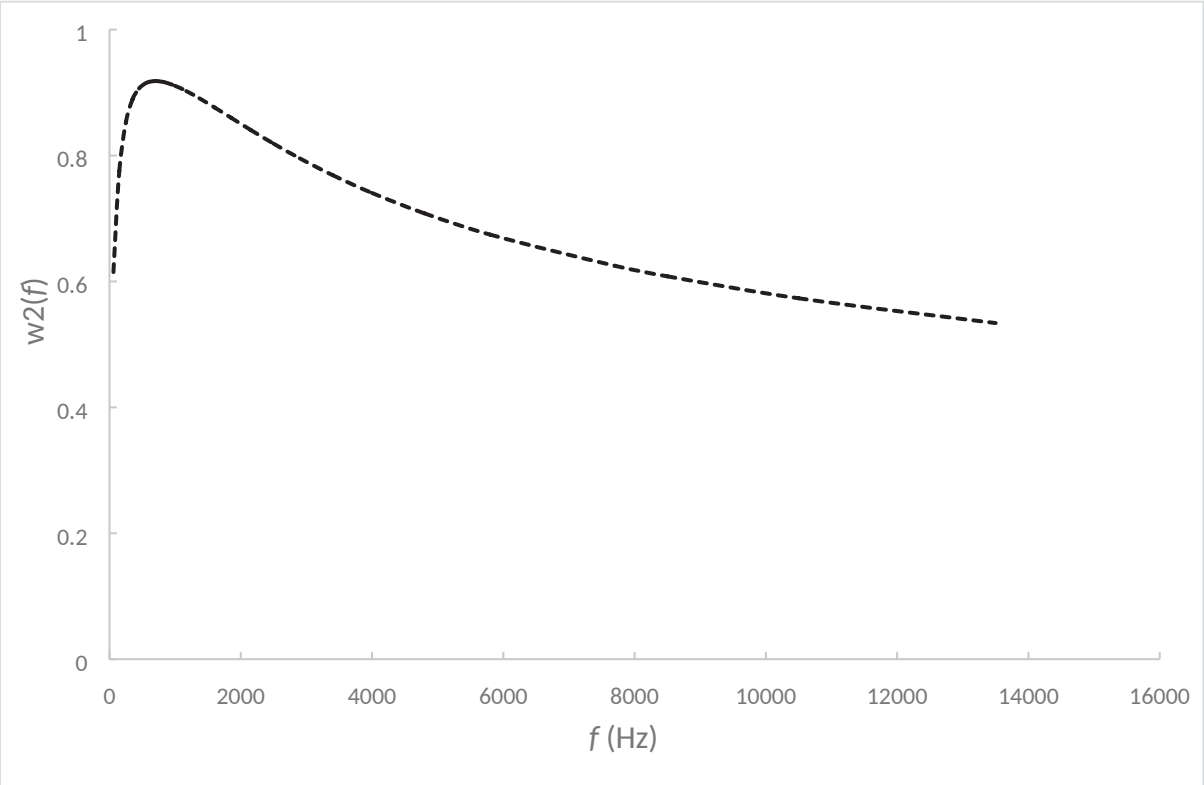


Figure 3. Weighting function $w_2(f)$ as a function of the frequency (f) in Hz of the detected tonal component (Eq. 3).

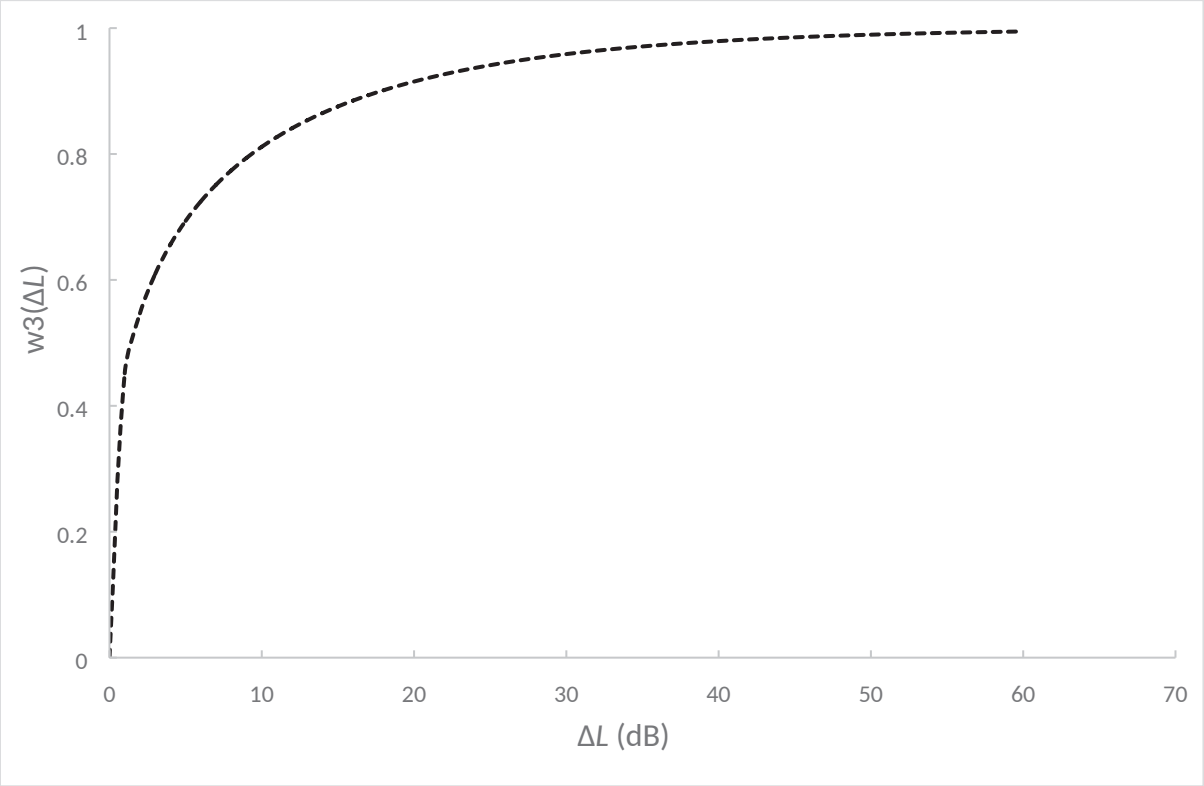


Figure 4. Weighting function $w_3(\Delta L)$ as a function of the level of the tonal component above the broadband masking noise (ΔL) (Eq. 4).

2.3. Sharpness

As proposed by von Bismark [37], Sharpness (measured in acum) is calculated using the specific loudness over critical band rate (N'), and a weighting function $g(z)$ which emphasized the high frequency content:

$$S = c\frac{\int_0^{24\ \mathrm{Bark}} N' g(z) dz}{\int_0^{24\ \mathrm{Bark}} N' dz} \quad (8)$$

Zwicker and Fast [35] defined $c = 0.11$, so a narrow-band noise centred at 1 kHz and bandwidth of 160 Hz with a sound-level of 60 dB produces a Sharpness of 1 acum, and also defined $g(z)$ as

$$g(z) = \begin{cases} 1 & \text{for } z \leq 16 \\ 0.066e^{0.171z} & \text{for } z > 16 \end{cases} \quad (9)$$

3. Material and methods

In order to achieve the research objectives, i.e. investigate whether the EPNL Tone Correction effectively assesses the human response to contemporary aircraft tonal noise, and whether Aures Tonality improves on the EPNL Tone Correction for such a purpose, a series of listening tests were carried out to gather data on human response to aircraft take-off noise (as briefly described in Section 1, and with further details in Section 3.1), to allow correlation with the psychoacoustic metrics tested in this research.

3.1. Stimuli

The stimuli used in the listening test were extracted from recordings of aircraft takeoffs made at a location approximately 900m from the end of the south runway of Heathrow airport (approx. 4.5 km from the Start-of-Roll (SOR) point) operating under westerly conditions. No information was available on the exact aircraft position during the measurement campaign, but using both the distance between the measurement and SOR points, and the standard flight profiles published by the Aircraft Noise and Performance (ANP) database[2] for each aircraft

[2] see https://www.aircraftnoisemodel.org/

recorded, the height of the aircraft passing over the measurement point is estimated as 435.2 ± 57.4 m. These recordings were made over 3 non-consecutive mornings, capturing all departing flight paths. Recordings were made using a class 1 Rion NL52 sound level meter with the microphone mounted (via a 5 m extension cable) on a 1 m diameter ground board in a relatively open position on flat ground. The microphone was covered with a hemispherical wind shield as shown in Fig. 5. The recordings were made in this way in order to minimise the influence of wind on the microphone, ground effects (i.e. constructive and destructive interference between the direct and reflected sound paths) and other extraneous noise sources. In this way it is considered that the fidelity of the recordings (in respect of the acoustic signature of the engine variants) has been maximised by reducing environmental factors to a minimum. Consecutive 1-hour long recordings were made, synchronised to the hour in order to enable identification of each take-off, using a sampling rate of 48 kHz and a 24 bit depth.

Figure 5. Noise recording setup.

The aircraft overflying the recording position were identified using live flight tracking information from uk.flightaware.com, allowing the make, model and engine to be assigned to each flyover time, which was noted during the recordings. From this list of aircraft and engines, the seven engine variants listed at Table 1were selected to be used in the listening tests.

Table 1

Data of the seven aircraft engines considered in the sound recording campaign.

Manufacturer	Engine variant	Entry into service date
General Electric	LEAP-(1 series)	2016
Pratt and Whitney	PW1127G	2015
Rolls Royce	Trent 1000	2011
General Electric	GE90-115B	2002
General Electric	GE90-(76B, 85B and 92B)	1995
Safran	CFM56-(5 series)	1994
Rolls Royce	RB211-(524 and 535)	1989

Table 2 shows the number and percentage of each different aircraft that was recorded. It should be noted that the A319, A320 and A321 aircraft are all very similar models used for short haul flights and that whilst these make up over 50% of the sample, it is likely that the percentage of these aircraft would be even higher if a larger sample was taken over an entire week (including all evening and early morning flights). From this list, a common engine variant (i.e. CFM56-5) used with the A320 family of aircraft was selected and for comparison with this, much more modern engine variants were selected (i.e. LEAP and PW1127G) which are used with the A320neo (the latest model produced by Airbus which is set to replace the

older A320 model). A similar selection process was carried out for the larger aircraft in the fleet whereby two older variants were selected (i.e. GE90-76/85/92B and RB211-5) to be representative of the older Boeing 757, 767 and 777 models and two newer variants (i.e. GE90-115B and TRENT1000) were selected to be representative of the newer Boeing 767, 777 and 787 models.

Table 2

Aircraft fleet sample recorded.

Aircraft	Count	Percentage
Airbus A320	120	29.56%
Airbus A319	68	16.75%
Boeing 777	64	15.76%
Boeing 767	25	6.16%
Airbus A321	25	6.16%
Boeing 787-9 Dreamliner	24	5.91%
Boeing 747	20	4.93%
Airbus A330	11	2.71%
Boeing 737	11	2.71%
Airbus A320 neo	10	2.46%
Airbus A380	5	1.23%
Boeing 757	5	1.23%
Boeing 787-8 Dreamliner	4	0.99%
DHC-8-402Q Dash 8	4	0.99%
Airbus A340	3	0.74%
Bombardier C-series	2	0.49%
Airbus A350	2	0.49%
Boeing 737 MAX 8	2	0.49%
British Aerospace 146-300A	1	0.25%

Although during the measurement campaign two and four engines aircraft were recorded, each aircraft that was included in the listening test had just two engines in order to avoid any perceptual differences introduced by the interaction of four engines compared with two. During the preparation of the RB211-5 variants stimuli subset to be used in the listening

tests, a noise sample of a Boeing 747-400 (four engines) was erroneously introduced. However, on the basis of the results of the listening tests (see Section 4, Fig. 9), the participants did not perceived the 747-400 noise sample differently to the other RB211-5 (two engine) noise samples.

Each instance of the engine variants shown in Table 1 were identified in the 1-hour recordings, and a total of 79 audio files, each 40 seconds in length were extracted, with the 40 seconds approximately centered on the time at which the aircraft was directly overhead. Any of these 40 s audio files which were found to contain a significant amount of extraneous background noise were discarded, and of the remaining a total of 59 files were selected for possible use in the test. Each of the selected files were auditioned further to find a 4 s clip which contained the most prominent audio character (usually just before the aircraft was overhead, but not always due to the comb filtering effect of ground interference varying the intensity of prominent audio characters). The most prominent audio character was generally perceived as the portion of the recording with the most prominent tonal content, but in some instances where tonal content was low this was perceived as the portion with the highest spectral variation between lower, mid and high frequencies. These 4 s clips were extracted and then a 0.5 s fade was applied to each end to create suitable comparable samples. During the inspection of the audio files, it was observed that the LEAP and PW1127G engine noise recordings had a significantly similar character, with the only variation being in the tonal content. Furthermore, the LEAP and PW1127G types are the engines of the A320neo. Therefore, it was decided to include the LEAP and PW1127G engines as the same variant (LEAP / PW1127G), thus the seven aircraft engine types shown in Table 1 became six aircraft engine variants for the purposes of the analysis. Finally, 48 4 s audio samples were selected for the listening test (8 per engine variant); also, another two 4 s audio samples were selected as reference for the engines of small aircraft (i.e. CFM56-5 and LEAP/PW1127G) and two

more 4 s audio samples for the engines of large aircraft (i.e. GE90-76/85/92B, GE90-115B, RB211-5 and Trent1000).

These (48 + 2 + 2) 4 s samples were the stimuli used in the listening tests, with the other seven potential 4 s samples discarded as being superfluous. Four seconds was selected as a suitable length of recording to allow the aircraft noise sample to be comprehended by the listener, but to be more easily comparable between noise samples. This also avoided too much pitch variation due to Doppler shift, and the fluctuation of complex tonal levels in comparison to broadband and BPF tone components within samples. Moreover, for the purposes of this research, the length of the stimuli and any variation within it was assumed not have any significant effect on subjective assessment [38-40]. It should be noted that complex tonal content emergence is generally close to zero from the point at which the plane is overhead due to the directionality of the noise sources involved.

All stimuli were normalised to an overall L_{Aeq} of 65 dB for specified gain settings with open-back headphones (using a B&K artificial ear coupling device) to minimise the effect of variations in recorded levels affecting preference, as otherwise 'quieter' recordings made of more distant aircraft would likely always be preferred, irrespective of the complex tonal content.

3.2. Experimental setup

The hardware setup used for the listening tests consisted of a mainstream laptop computer with an Audiotest DragonFly Red USB DAC/Headphone Amplifier and a pair of AKG k-501 open-back headphones. The tests were carried out in a very quiet environment (i.e. an audiology room of the Institute of Sound and Vibration Research at the University of Southampton), with no interference from outside in order to avoid distractions.

The test was entirely automated via a bespoke MatLab code. The volume level on the laptop was always set to maximum, with MatLab controlling the playback volume to ensure consistency.

3.3. Participants

The listening tests were undertaken by 35 healthy participants (25 males and 10 females). The average age of the participants was 30.5 ± 9.2 years old (57% between 20-29 years old, 31% between 30-39 years old, 6% between 40-49 years old, and 6% between 50-59 years old). A thank you gift of £10 for taking part was used to incentivize participation in the listening tests. Prior to participating in the listening test, each participant was required to confirm normal hearing ability[3] and asked to fill out a consent form. This experiment was approved by the Ethics and Research committee of the University of Southampton.

3.4. Experimental procedure

The listening tests involved a series of identical tasks, asking the participants to rank sets of six 4 s stimuli in order of preference. The participants were required to rank each of the six 4 s stimuli from most preferable to the least preferable (see Fig. 6). During the process of ranking by order of preference, participants were allowed to listen to each individual stimulus as many times as they required until the final order was decided. Once the order of preference was decided, the participants were required to listen to all the samples in order of preference. After listening to all the (six 4 s) samples in order of preference, the participants had the

[3] In case they felt any doubt about this requirement, the participant was directed to the Action on Hearing Loss charity telephone test on 0844 800 3838 or online test: http://www.actiononhearingloss.org.uk/your-hearing/look-after-your-hearing/check-your-hearing/take-the-check.aspx.

opportunity to change their response until they were satisfied. Once the final order of preference was confirmed they were able to continue with the test. The same process was used for the remaining sets of six 4 s stimuli. It should be noted that no specific order of presentation was suggested to the participants, so they could start listening to the six 4 s stimuli in the order they wanted, for instance, from stimulus A to stimulus F, from stimulus F to stimulus A, etc. (see Fig. 6).

The stimuli sets each comprised four samples, and two reference samples which were included in each set (making six samples in total per set) (see Fig. 7). In total 72 stimuli were ordered by each participant (including 24 instances of one of the four reference samples being ordered) in three groups of four sets each. The first group of sets was made up of the two variants of smaller engines (CFM56-5 and LEAP / PW1127G), whilst the second and third groups were a mix of the four variants of larger engines (RB211-5, GE90-76/85/92B, GE90-115B and TRENT1000). The small and large engines differ significantly in their dimensions and thrust produced, and therefore, remarkably different noise signatures between them were anticipated. For this reason, different reference samples (selected from samples of the engine variants presented in the group) were used for the small engines in the first group and the large engines in the second and third groups. It should therefore be noted that the calculated preference magnitude of the small and large engine variants is not directly comparable. However, since the aim of this study is to look at metrics which correlate with preference magnitude for each of the small and large engine variants, this does not represent any bias.

Without considering the reference samples (2 x 4 sets of stimuli x 3 groups = 24 reference samples), the participants evaluated 48 stimuli, i.e. 8 stimuli for each of the 6 engine variants. Each of these 48 individual stimuli was only once presented to the participant, although as described above, they could listen to each of them as many times as required until the final order of preference was confirmed. After each of the three groups of four sets, the

participants were required to describe (with their own words) the audio samples ranked as the least and most preferable in the first and fourth sets, and to describe how one compared to the other. Note that this information has not been used in this paper. In order to minimize the listener's fatigue as far as possible, the participants were instructed to pause whenever they needed to stay focus on the task required. Overall, the participants required between 50 and 75 min to complete the listening test.

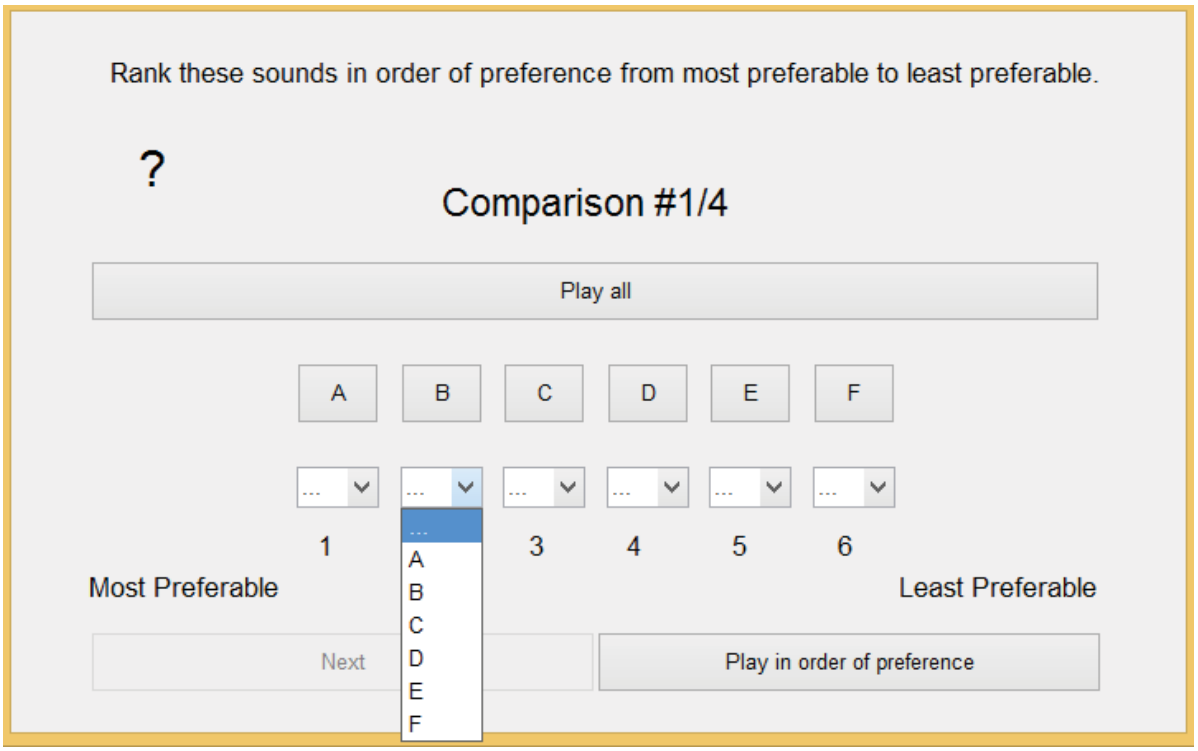


Figure 6. Example of the interface used by the participants during the listening tests.

The two reference samples used in each set of stimuli were selected to be two different engine models in order to ensure differing spectral content and maximise the difference in their likely perceptual qualities compared to each other. As described above, two stimuli were used as reference in the first group of sets, and another two reference stimuli selected for the second

and third groups of sets (the same two for both of these groups). The two reference samples were selected on the basis of the criteria and expertise of the researchers involved in this research, after analysing the outcomes of pilot studies conducted prior to the development of the final formal listening tests described in this paper. The decision to use a preference ordering method, as opposed to a scoring method, was taken to avoid the inherent variation between participants interpretation of a scoring system (where a number between X and Y is assigned for preference or annoyance) where, often, some participants will assign a limited range of values to a dataset where other participants will try and use the full allowable range for the same dataset. It is also considered that comparative ordering of preference is an easier task for the average listener than assigning a number to a single sample [41].

A preference ordering procedure with two reference samples, trialled and developed by the authors, is used for calculating a preference magnitude (in an interval scale) for each of the 48 audio samples evaluated. After the analysis of the results of previous pilot studies, the authors observed that a two reference preference ordering method was more suitable than a one reference method for the purposes of this study, as it allows a more dynamic assignment of preference for each sample in relation to the reference samples (i.e. preference magnitude depends not only on the ranking of each sample but the relative spacing of the reference samples). This increased dynamic range of the preference magnitude allows for potentially more representative values that are more readily compared between each listening test participant. Fig. 7 shows the process for magnitude assignment (hereinafter referred to as Preference Rating (PR)) based on the stimuli order by preference (with two reference stimuli). As can be seen in Fig. 7, an arbitrary magnitude of 60 and 40 was set for the reference samples 1 and 2 (note that the magnitudes of 40 and 60 would switch between reference 1 and 2 if the order of these samples was reversed). Depending on the order of these two references within

the set, the magnitude (or PR) for each sample is assigned on the basis of equal spacing using the arbitrary magnitude of the reference samples.

Sample Order	Assigned Magnitude
Reference 1	60.0
Sample A	56.0
Sample B	52.0
Sample C	48.0
Sample D	44.0
Reference 2	40.0
Magnitude Range	100%

Sample Order	Assigned Magnitude
Sample A	65.0
Reference 1	60.0
Sample B	55.0
Sample C	50.0
Sample D	45.0
Reference 2	40.0
Magnitude Range	125%

Sample Order	Assigned Magnitude
Sample A	66.7
Reference 1	60.0
Sample B	53.3
Sample C	46.7
Reference 2	40.0
Sample D	33.3
Magnitude Range	167%

Sample Order	Assigned Magnitude
Sample A	70
Reference 1	60
Sample B	50
Reference 2	40
Sample C	30
Sample D	20
Magnitude Range	250%

Sample Order	Assigned Magnitude
Sample A	100
Sample B	80
Reference 1	60
Reference 2	40
Sample C	20
Sample D	0
Magnitude Range	500%

Figure 7. Illustration of the possible outcomes of assigned magnitude based on the position of the two reference stimuli in the order of preference.

4. Results

4.1. Preference Rating

Table 3 shows the average Coefficient of Variation (CV), calculated as the standard deviation divided by the mean, Interquartile Range (IQR), calculated as Quartile 3 (Q3) minus Quartile 1 (Q1), and the number of extreme outliers of the participants' PR values for each variant of aircraft engines. A participant's PR beyond the lower fence = Q1 - 3×IQR or the upper fence = Q3 + 3×IQR is considered an extreme outlier. As can be seen in Table 3, the

inter-participant variability is similar among all the variants of aircraft engines tested. The highest CV is for the RB211-5 engine variant. Observing the IQR, most of the participants' PR values for the GE90-115B, LEAP / PW1127G and TRENT1000 engine variants are concentrated around the median values. Four and three outliers are identified in the participants' responses for the LEAP and TRENT engine variants respectively. The presence of these outliers has a very minor influence on the calculated PR values; i.e. a relative deviation of 1.89% (LEAP/PW1127G variant) and 0.68% (TRENT1000 variant) for the PR values calculated with and without outliers.

Table 3

Average coefficient of variation (CV) and interquartile range (IQR), and number of outliers ($> 3 \times$ IQR) of the participants' PR values for each aircraft engine variant.

Engine variants	Coefficient of Variation	Interquartile Range	Outliers
CFM-56-5	0.45	41.66	0
GE90-76/85/92B	0.43	42.50	0
GE90-115B	0.51	30.20	0
LEAP / PW1127G	0.51	22.08	4
RB211-5	0.54	42.75	0
TRENT1000	0.43	31.20	3

A certain degree of variability in the participants' responses is expected in experiments involving the perceptual assessment of noise stimuli [42]. As explained above, during the listening test each participant created their own individual scale (for each set of stimuli) on the basis of the relative position of the two reference stimuli, and therefore, some inter-participant

variability was expected. To measure the agreement among the different participants when ranking the series of stimuli presented during the listening tests, a (non-parametric) k-related samples statistic, Kendall's W [43 (Section 3)], was calculated for each variant of aircraft engines. For each variant of aircraft engines, the null hypothesis that there was no agreement among the rank orderings given by the participants (i.e. W=0) was tested. In order to ensure a robust calculation of the p-values, a Monte Carlo bootstrapping with 10000 samples was applied. As shown in Table 4, the p-values are smaller than 0.05 for all the groups of aircraft engines, thereby allowing rejection of the null hypothesis (of no agreement between participants' rankings) for all engine variants.

Table 4

Results of the Kendall's W statistic for each variant of aircraft engines tests. *p-value calculated with a Monte Carlo bootstrapping with 10000 samples.

	CFM-56-5	GE90-76/85/92B	GE90-115B	LEAP / PW1127G	RB211-5	TRENT1000
Kendall's W	0.205	0.057	0.092	0.381	0.064	0.131
p-value	0.000	0.048	0.002	0.000	0.028	0.000

It is well known that age alters the hearing of high frequencies, and therefore can alter the perception of tones in the high frequency region. To test whether there are differences in the PR values across the different age intervals (see sub-section 3.3), Mood's median test of

independent samples[4] [44] were carried out. Based on the results of these statistical tests, at a significance level of 0.05, the null hypothesis of equal PR values across the different age intervals can be retained: first group of stimuli sets (p-value = 0.559), second and third groups of stimuli sets (p-value = 0.168).

4.2. Acoustics characterisation of the aircraft engine variants tested

Mood's median tests of independent samples[4] [44] were conducted for testing the null hypothesis that the medians of the PNL, EPNL Tone Correction, Aures Tonality and Sharpness metrics are the same across engine variants of the small aircraft (CFM-56-5 and LEAP/PW1127G), and across engine variants of the large aircraft (GE90-76/85/92B, GE90-115B, RB211-5 and TRENT1000). For the engines of small aircraft, statistically significant differences, at a significance level of 0.05, are observed between the medians of the PNL (p-value = 0.010) and Sharpness (p-value = 0.010) metrics of the CFM-56-5 and LEAP/PW1127G variants. As shown in Table 5, statistically significant differences (at a significance level of 0.05) are found between the four engine variants with each other in the medians of at least one of the noise metrics used in this paper. The results of this statistical analysis support the assumptions that, in terms of the noise metrics used in this paper, the two engine variants of small aircraft are different from each other, and that the four engine variants of large aircraft are different from each other.

[4] Due to the small number of samples, and that the data do not follow a normal distribution (Shapiro-Wilk test), a non-parametric test was selected.

Table 5

Noise metrics with medians with statistically significant differences ($p \leq 0.05$) between each pair of the 4 engine variants of large aircraft compared, based on the results of the Mood's median test of independent samples.

	GE90-76/85/92B	GE90-115B	RB211-5	TRENT1000
GE90-76/85/92B	-	PNL (p-value = 0.016), EPNL Tone Correction (p-value = 0.016), Aures Tonality (p-value = 0.016), Sharpness (p-value = 0.016)	PNL (p-value = 0.016)	PNL (p-value = 0.000), EPNL Tone Correction (p-value = 0.016), Aures Tonality (p-value = 0.000), Sharpness (p-value = 0.016)
GE90-115B		-	Sharpness (p-value = 0.016)	PNL (p-value = 0.000)
RB211-5			-	Sharpness (p-value = 0.016)
TRENT1000				-

Although all the stimuli were normalized to 65 dBA, there is a range of up to 3 PNdB in the average Perceived Noise Level (PNL) between the different aircraft engine variants (as shown at Table 6). This finding is in line with previous studies suggesting that metrics based on dBA alone are not able to appropriately explain perceived annoyance caused by aircraft noise [17-18].

Moreover, as seen in Table 6, there are notable differences in the average EPNL Tone Correction and Aures Tonality among the aircraft engine variants tested, with the GE90-115B and TRENT1000 engine variants having the lowest tonality as assessed using both metrics. The average values of Sharpness differ also among the different aircraft engine variants, with the LEAP/PW1127G engine variant having the lowest value of Sharpness. Further differences in the frequency spectra (and especially in the tonal content) between a representative sample of each engine variant are described below.

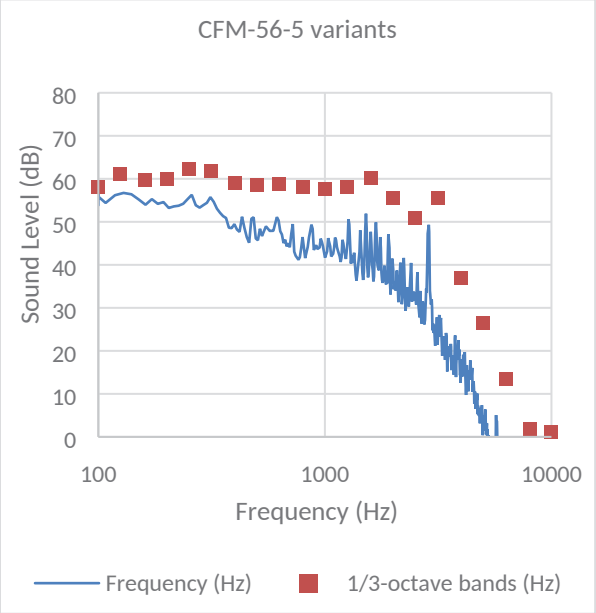
Table 6

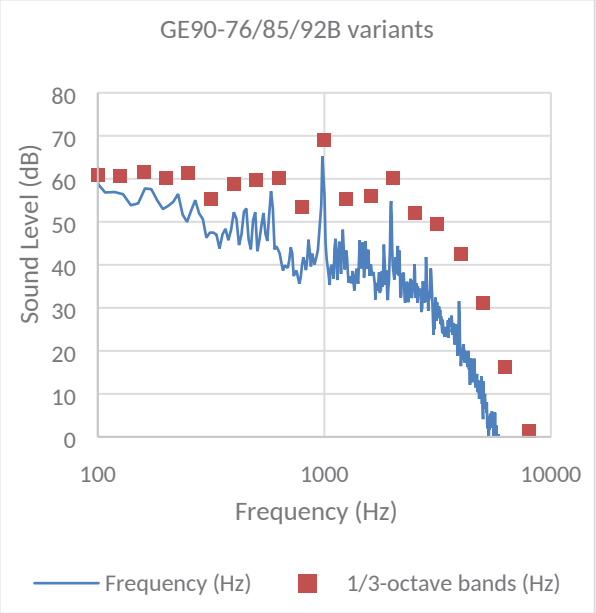
Average value (and standard deviation) of the Perceived Noise Level (PNL), EPNL tone correction, Aures tonality and Sharpness metrics for the aircraft engine variants assessed.

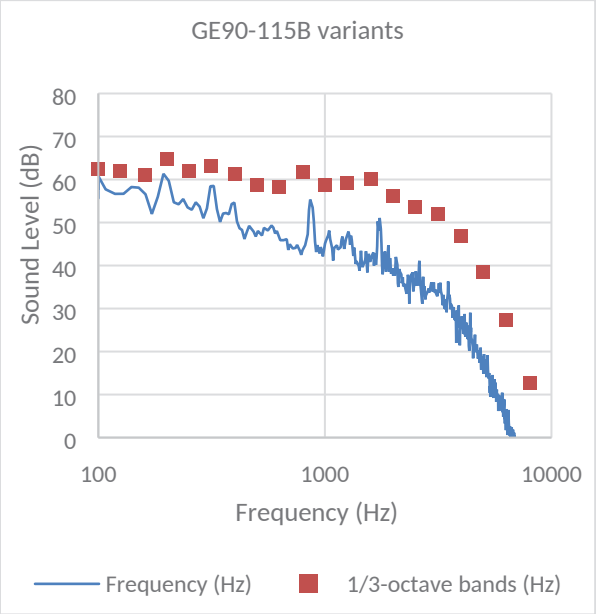
Engine variants	PNL (PNdB)	EPNL Tone Correction (dB)	Aures Tonality (tu)	Sharpness (acum)
CFM-56-5	88.30±0.86	2.10±0.61	0.30±0.07	1.98±0.24
GE90-76/85/92B	87.44±0.63	3.43±1.08	0.30±0.06	1.81±0.11
GE90-115B	88.76±0.25	1.21±0.43	0.16±0.04	2.06±0.11

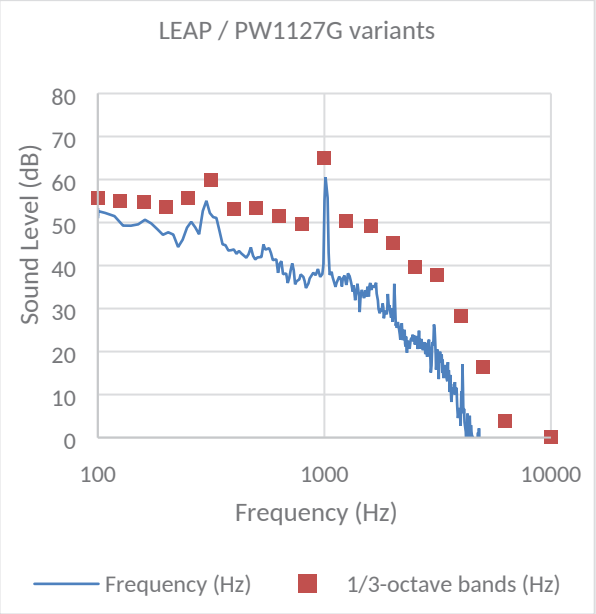
LEAP / PW1127G	86.81±0.92	2.66±1.57	0.23±0.07	1.53±0.13
RB211-5	88.80±0.66	1.54±0.82	0.26±0.08	1.86±0.13
TRENT1000	89.82±0.51	1.44±0.46	0.17±0.02	2.15±0.15

Fig. 8 shows frequency spectra and 1/3-octave band sound-levels for a set of noise samples which highlight the differences between the engine variants. The frequency spectra and 1/3-octave band sound-levels were calculated as the average of the central 0.5 s of the 4 s aircraft noise samples. The samples presented at Fig. 8 are those stimuli reported as least preferable, i.e. lowest PR for each engine variant. The CFM-56-5 sample (Fig. 8 – top left) has a high-frequency blade passing frequency (BPF) tone and clearly discernible complex tones in the mid-to-high frequency region. The GE90-76/85/92B sample (Fig. 8 – top right) has a clearly identifiable mid-frequency BPF tone (and harmonics at higher frequencies), and also clearly discernible complex tones spread across the spectrum. The GE90-115B sample (Fig. 8 – middle left) is similar to the GE90-76/85/92B sample, with a mid-frequency BPF tone, albeit with a lower prominence, and complex tones spread across the spectrum, which are less clearly identified due to masking by broadband noise. The LEAP / PW1127G sample (Fig. 8 – middle right) has a very clearly identifiable mid-frequency BPF tone and harmonics at higher frequencies. In the RB211-5 sample (Fig. 8 – bottom left) a series of complex tones with relatively similar sound-levels are clearly discernible in the low-to-mid frequency region, with no prominent BPF tone. Finally, the TRENT1000 sample (Fig. 8 – bottom right) has a significant content in low-frequency broadband noise and a series of discrete (possibly harmonic) spaced tones spread across the spectrum.


CFM-56-5 variants
Sound Level (dB)
80
70
60
50
40
30
20
10
0
100
1000
10000
Frequency (Hz)
Frequency (Hz)
1/3-octave bands (Hz)


GE90-76/85/92B variants
Sound Level (dB)
80
70
60
50
40
30
20
10
0
100
1000
10000
Frequency (Hz)
Frequency (Hz)
1/3-octave bands (Hz)


GE90-115B variants
Sound Level (dB)
80
70
60
50
40
30
20
10
0
100
1000
10000
Frequency (Hz)
Frequency (Hz)
1/3-octave bands (Hz)


LEAP / PW1127G variants
Sound Level (dB)
80
70
60
50
40
30
20
10
0
100
1000
10000
Frequency (Hz)
Frequency (Hz)
1/3-octave bands (Hz)

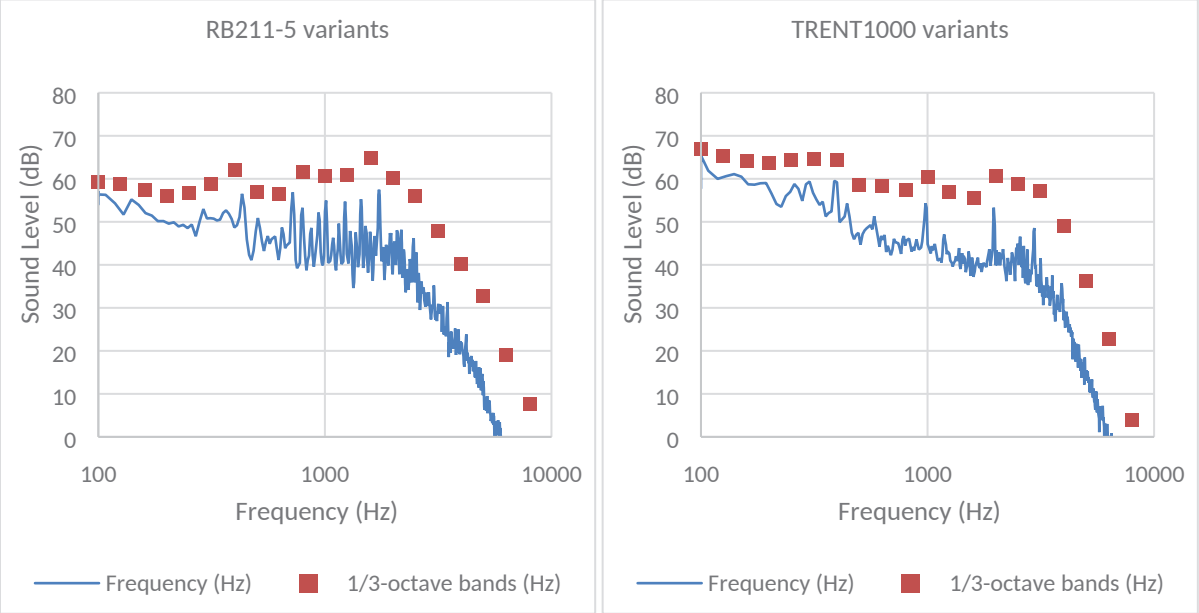


Figure 8. Frequency spectrum and 1/3-octave band sound-levels of the least preferable sample of aircraft engine variants.

4.3. EPNL Tone Correction and Aures Tonality vs. Preference rating

Although the purpose of this paper is not to compare differences in perception between each engine variant, but to analyse whether the noise metrics used can explain the differences in perception within each engine variant, a related-samples Wilcoxon Signed Rank test[4] [44] was performed to test the null hypothesis that the median of differences in PR between the CFM-56-5 and LEAP/PW1127G variants equals 0, and multiple related-samples Wilcoxon Signed Rank tests were performed to test the null hypothesis that the median of differences in PR between the GE90-76/85/92B, GE90-115B, RB211-5 and TRENT1000 variants with each other equals 0. The distinction between engines of small and large aircraft for this statistical analysis is due to, as described in sub-section 3.4, the used of different reference samples for the first group (small aircraft) and second/third groups (large aircraft) of stimuli.

For the engines of the small aircraft, the null hypothesis can be rejected (at a significant level of 0.05), and therefore, statistically significant differences in PR are observed between

the CFM-56-5 and LEAP/PW1127G variants (p-value = 0.000). For the engines of the large aircraft, the null hypothesis of median of the differences in PR equals to 0 can only be rejected (at a significance level of 0.05), and therefore, statistically significant differences in PR are only observed for the pair-comparisons: GE90-76/85/92B – GE90-115B (p-value = 0.001), GE90-76/85/92B – RB211-5 (p-value = 0.005), GE90-115B – RB211-5 (p-value = 0.029) and GE90-76/85/92B – TRENT1000 (p-value = 0.001). The null hypothesis cannot be rejected for the pair-comparisons: GE90-76/85/92B – TRENT1000 (p-value = 0.064) and RB211-5 – TRENT1000 (p-value = 0.781). If a conservative approach is taken, and a Bonferroni correction [44] is applied for multiple comparisons, then the null hypothesis should not be rejected for the pair-comparison: GE90-115B – RB211-5 (Bonferroni corrected p-value = 0.174). These results should, however, be interpreted with caution since, as mentioned above, this experiment was designed for assessing differences in perception (i.e. PR) within each engine variant rather than between each engine variant.

The performance of Aures Tonality and EPNL Tone Correction for assessing the perception of tonal content in aircraft noise was investigated, using six variants of aircraft engines of diverse EIS dates and of aircraft from narrow to wide body. A series of multiple linear regression (MLR) analyses were carried out using the PR (calculated from the participants' responses as describe above) as dependent variable and (i) PNL (Models M.0); (ii) PNL and EPNL Tone Correction (Models M.1); and (iii) PNL and Aures Tonality (Models M.2) as independent variables.

As shown in Table 7, and corresponding with findings by several researchers [19,30,45], the EPNL Tone Correction is not always able to account for the perceived effect of the tonal content of the aircraft noise samples on the preference reported by the participants (i.e. PR). As expected the EPNL Tone Correction is only able to explain the variance of the PR, as obtained from the participants' responses, for aircraft noise samples with the tonal

content composed solely of a physically dominant BPF tone (LEAP / PW1127G variants) (R^2 = 0.76).

Table 7

Multiple Linear Regression (MLR) results with PNL, EPNL Tone Correction and Aures Tonality as independent variables, and PR as dependent variable for each aircraft engine variant.

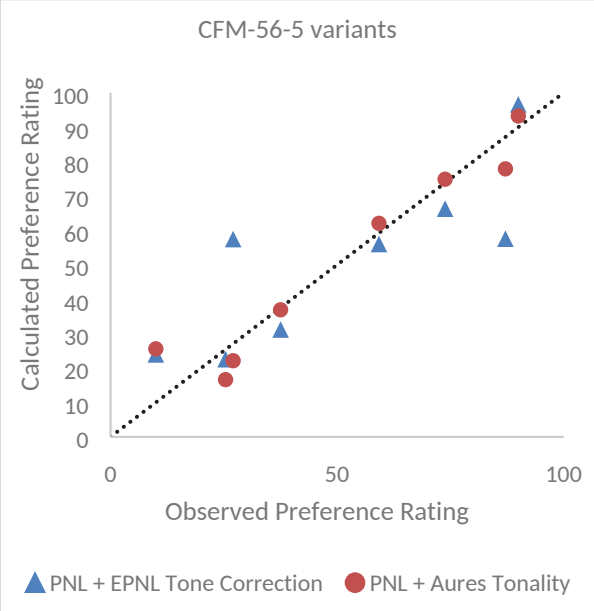
Engine variants	Model	Independent variables	R	R-square	Adjusted R-square	Std. error of the estimate	F-value
CFM-56-5	M.0	PNL	0.81	0.66	0.61	7.04	11.82
	M.1	PNL + EPNL tone correction	0.82	0.67	0.54	7.65	5.06
	M.2	PNL + Aures tonality	0.97	0.93	0.90	3.49	33.70
GE90-76/85/92B	M.0	PNL	0.54	0.29	0.18	6.39	2.49
	M.1	PNL + EPNL tone correction	0.61	0.37	0.12	6.61	1.47
	M.2	PNL + Aures tonality	0.54	0.29	0.01	7.00	1.04
GE90-115B	M.0	PNL	0.60	0.36	0.26	6.61	3.41
	M.1	PNL + EPNL tone correction	0.66	0.44	0.21	6.81	1.93

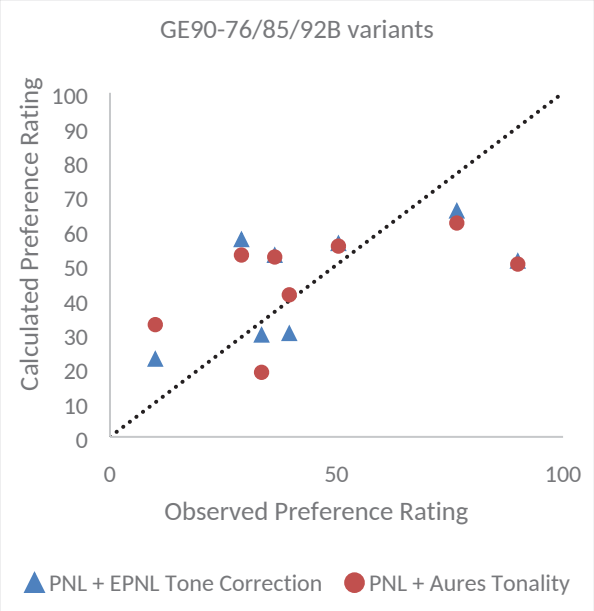
	M.2	PNL + Aures tonality	0.87	0.76	0.66	4.46	7.81
LEAP / PW1127G	M.0	PNL	0.32	0.10	-0.05	14.42	0.69
	M.1	PNL + EPNL tone correction	0.87	0.76	0.66	8.25	7.72
	M.2	PNL + Aures tonality	0.89	0.78	0.70	7.75	9.07
RB211-5	M.0	PNL	0.46	0.21	0.08	8.10	1.57
	M.1	PNL + EPNL tone correction	0.61	0.37	0.12	7.91	1.47
	M.2	PNL + Aures tonality	0.84	0.71	0.59	5.37	6.11
TRENT1000	M.0	PNL	0.95	0.90	0.88	3.06	51.19
	M.1	PNL + EPNL tone correction	0.96	0.93	0.90	2.78	32.29
	M.2	PNL + Aures tonality	0.96	0.91	0.88	3.06	26.12

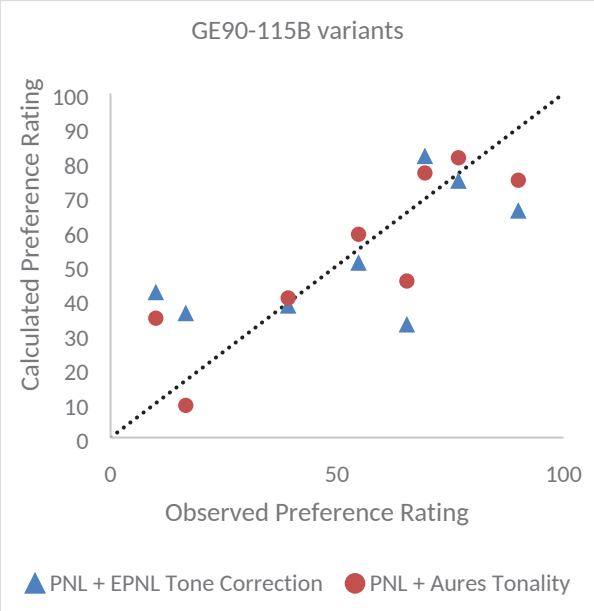
As seen in Table 7, Aures Tonality notably outperforms the EPNL Tone Correction when explaining the variance in the participants' responses in terms of PR. The results obtained in this research are in line with the assumption that Aures Tonality improves on the EPNL Tone Correction in terms of accounting for the presence of complex tones in aircraft noise [46]. For the aircraft noise samples with a significant content in complex tones, i.e. CFM-56-5, GE90-115B and RB211-5 engine variants, the R^2 coefficients for model M.2 (including Aures

Tonality), are 0.93, 0.76 and 0.71 respectively which show significant improvement over the R^2 coefficients for Model M.1 (0.67, 0.44 and 0.37 respectively). Even for the aircraft noise samples with a physically dominant BPF tone (LEAP / PW1127G), Aures Tonality has a slightly better correlation with the PR compared to EPNL Tone Correction, as Aures Tonality might be able to capture the contribution of the BPF harmonics. There are two exceptions: (1) the TRENT1000 engine noise samples, where the main contributor to the PR was the loudness (i.e. PNL), and the tonal content does not seem to have made a significant contribution to the participants' responses; (2) the GE90-76/85/92B engine noise samples where there is a particular feature influencing the participants' responses (see Section 4.4) which is not appropriately captured by any of the noise metrics evaluated, i.e. PNL, EPNL Tone Correction and Aures Tonality.

The observed and calculated PR, using PNL plus EPNL Tone Correction (blue triangles) and PNL plus Aures Tonality (red circles) is shown in Fig. 9 for the six aircraft engine variants tested. In Fig. 9 it can be seen that the PR calculated using the PNL plus the EPNL Tone Correction has a worse correlation with the observed PR than the one calculated with PNL plus Aures Tonality, for all the engine variants tested but for the two exceptions noted above (TRENT1000 and GE90-76/85/92B engines).


CFM-56-5 variants
Calculated Preference Rating
Observed Preference Rating
0
10
20
30
40
50
60
70
80
90
100
0
50
100
PNL + EPNL Tone Correction
PNL + Aures Tonality


GE90-76/85/92B variants
Calculated Preference Rating
Observed Preference Rating
0
10
20
30
40
50
60
70
80
90
100
0
50
100
PNL + EPNL Tone Correction
PNL + Aures Tonality


GE90-115B variants
Calculated Preference Rating
Observed Preference Rating
0
10
20
30
40
50
60
70
80
90
100
0
50
100
PNL + EPNL Tone Correction
PNL + Aures Tonality

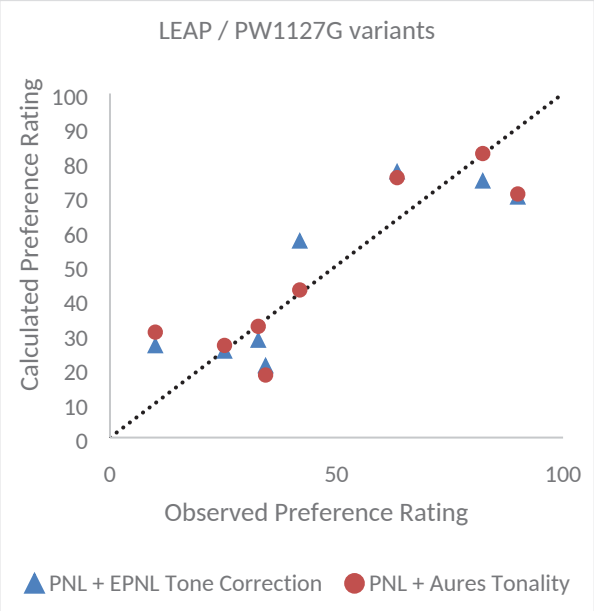

LEAP / PW1127G variants
Calculated Preference Rating
Observed Preference Rating
0
10
20
30
40
50
60
70
80
90
100
0
50
100
PNL + EPNL Tone Correction
PNL + Aures Tonality

Figure 9. Observed vs. Calculated Preference Rating (PR) for each aircraft engine variant. *Observed PR normalised to the range [10-90], with 10 = 'least preferable' and 90 'most preferable'.

4.4. Effect of Sharpness on Preference Rating

During a series of interviews after the listening test was finished, the participants reported the amplitude modulated noise (which the authors attributed to the BSN present in most of the aircraft noise samples tested), and also the 'high pitch' noise as the least preferable noise features in the aircraft noise they listened to. There is no agreement in the specific order of these two noise features, as some participants reported the amplitude modulated noise as the least preferable feature, and others the 'high pitch' noise. 'High pitch' noise has been found to be an important contributor to the human perception of aircraft noise during takeoff [47]. However, Gille et al. [48] found that Sharpness metric did not correlate with annoyance caused by aircraft sounds during takeoff, although the participants reported the aircraft sounds as 'sharp' (as a consequence of the emergence of high frequencies).

A series of MLR analyses are carried out with Sharpness along with PNL and Aures Tonality as independent variables (Models M.3), and the PR as dependent variable. As shown in Table 8, the inclusion of Sharpness only plays an important part in explaining the variance of the PR for two engine variants, the GE90-76/85/92B ($R^2 = 0.76$) and the GE90-115B ($R^2 = 0.86$). As described above, neither the PNL alone, nor it combined with the tonality methods tested (EPNL Tone Correction and Aures Tonality) are sufficient to explain the variance of the PR for the GE90-76/85/92B noise samples. The results shown in Table 8 indicate, however, that the PR for the GE90-76/85/92B noise samples is well correlated when the changes in Sharpness are taken into account in addition. Moreover, for the GE90-115B noise samples, there is an increase in the correlation with the PR when Sharpness is added to PNL and Aures Tonality as independent variable (Models M.3). However, no effect of Sharpness on PR is observed for the other four aircraft engine variants. As mentioned above, the changes of PR in these other four aircraft engine variants is driven by the variations of tonal content: BSN (CFM56-5 and RB211-5 variants) and BPF tone (LEAP/PW1127G variant); and loudness (TRENT1000). In line with these results, Barbot et al. [47] found that Sharpness did not correlate with the perception of aircraft noise when the presence of BSN was the dominant feature.

Table 8

MLR results with PNL, Aures tonality and Sharpness as independent variables, and PR as dependent variable for each aircraft engine variant.

Engine variants	Model	Independent variables	R	R-square	Adjusted R-square	Std. error of the estimate	F-value

CFM-56-5	M.2	PNL + Aures tonality	0.97	0.93	0.90	3.49	33.70
	M.3	PNL + Aures tonality + Sharpness	0.97	0.93	0.88	3.87	18.29
GE90-76/85/92B	M.2	PNL + Aures tonality	0.54	0.29	0.01	7.00	1.04
	M.3	PNL + Aures tonality + Sharpness	0.87	0.76	0.58	4.56	4.21
GE90-115B	M.2	PNL + Aures tonality	0.87	0.76	0.66	4.46	7.81
	M.3	PNL + Aures tonality + Sharpness	0.93	0.86	0.76	3.73	8.50
LEAP / PW1127G	M.2	PNL + Aures tonality	0.89	0.78	0.70	7.75	9.07
	M.3	PNL + Aures tonality + Sharpness	0.89	0.79	0.63	8.62	4.90
RB211-5	M.2	PNL + Aures tonality	0.84	0.71	0.59	5.37	6.11
	M.3	PNL + Aures tonality + Sharpness	0.85	0.72	0.51	5.88	3.45

TRENT1000	M.2	PNL + Aures tonality	0.96	0.91	0.88	3.06	26.12
	M.3	PNL + Aures tonality + Sharpness	0.96	0.93	0.87	3.17	16.49

Fig. 10 shows the observed and calculated PR with the PNL plus Aures Tonality (filled red circles) and with the PNL plus Aures Tonality plus Sharpness (unfilled green circles). As seen in Fig. 10, for the GE90-76/85/92B (Fig 10 – left) and GE90-115B (Fig. 10 – right) noise samples the inclusion of Sharpness in the MLR allows a better correlation between the calculated and observed PR.

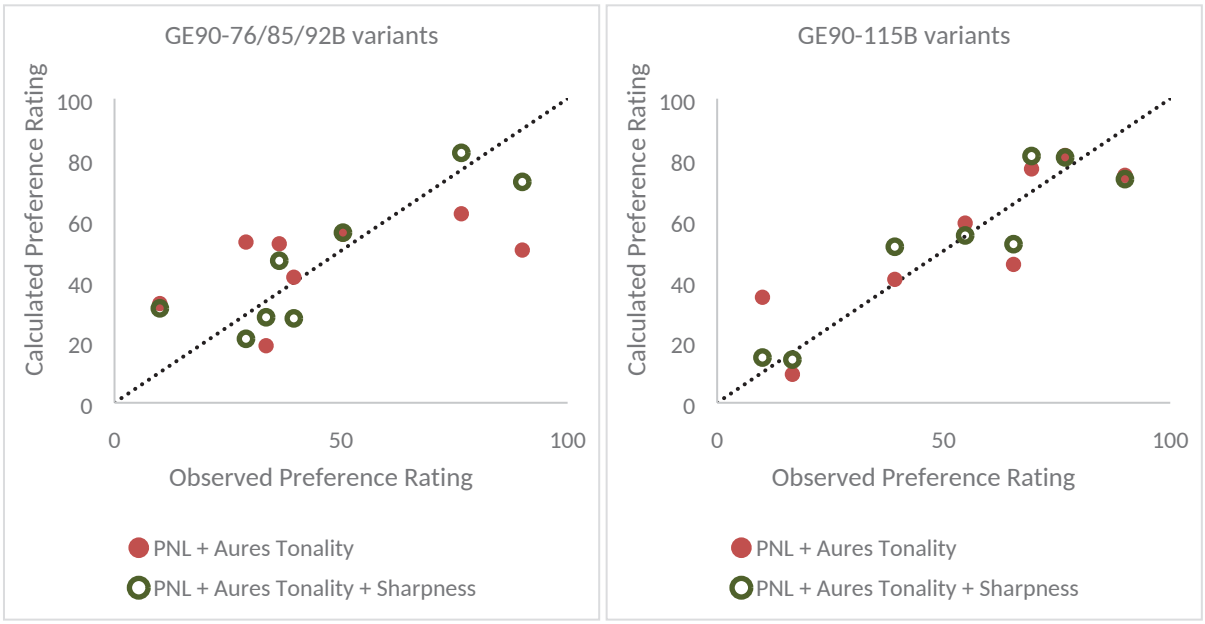


Figure 10. Observed vs. Calculated Preference Rating (PR) for GE90-76/85/92B and GE90-115B engine variants. *Observed PR normalised to the range [10-90], with 10 = 'least preferable' and 90 'most preferable'.

5. Discussion

The EPNL Tone Correction is unable to account for the perceived effect of the tonal content of most of the aircraft noise samples tested in this research, which are considered to be a representative sample of the contemporary aircraft fleet using the UK's busiest airport. Observing the Adjusted R^2 and F-values shown in Table 7, the addition of the EPNL Tone Correction to the PNL as a predictor of the PR (Models M.1) only provides a better fit to the PR data, as compared to models with only PNL as predictor (Models M.0), for the aircraft engine variant with the tonal content composed solely of a physically dominant BPF tone (LEAP / PW1127G engines). The EPNL Tone Correction is based on the magnitude of the strongest protruding tone [14], and therefore is not able to account for the perceptual effect of the harmonics of that physically dominant tone, or other complex tones that may be present. Moreover, the EPNL Tone Correction uses a simplistic frequency division, i.e. 1/3-octave bands, for identifying and calculating tonal penalties [30], and, as illustrated in Fig. 8, this is an insufficient resolution to be able to detect the presence of a series of complex tones.

Aures Tonality improves on the EPNL Tone Correction in terms of correlation with the subjective response to the aircraft noise samples that contain diverse tonal content tested in this research. Aures Tonality is able to account for the effect of a physically dominant tone (such as BPF tone) and its harmonics, and a series of complex tones on the PR obtained from the participants' responses. When added to the PNL as a predictor of the PR, Aures Tonality (Models M.2) provides a significantly better fit to the PR data, as compared to models with only PNL as predictor (Models M.0) (see the Adjusted R^2 and F-values shown in Table 7), for all the aircraft engine variants but for the two exceptions indicated above, i.e. GE90-76/85/92B and TRENT1000 (see below for further explanation). These results are in line with Minard et al. [46], where Aures Tonality was found as the most accurate tonality metric for assessing the perceived unpleasantness of a series of synthesized aircraft sounds during takeoff. Minard et

al. [46] found Aures Tonality to be able to detect and account for the unpleasantness of complex tones, as the weighting function of the tonal prominence over the broadband masking noise (Eq. 4) is calculated over the whole spectrum and considering both isolated tones and complex tones covering a wide frequency band.

Notwithstanding the significant improvement of Aures Tonality over the EPNL Tone Correction when accounting for the perceptual effect of the tonal content of the aircraft engine noise samples tested, as seen in Fig. 9, there are some calculated values of PR that differ remarkably from the average observed values. The latter is especially apparent in aircraft noise samples with significant complex tonal content (e.g. RB211-5 variant). These aircraft noise samples contain a series of complex tones spaced evenly across the frequency spectrum with relatively even sound levels. Although each complex tone considered individually might not be saliently perceived alone, it has been suggested that the interaction between complex tones plays an important part in the perception of aircraft noise [46]. The procedure for calculating Aures Tonality (Section 2.2) does not include a function to take into account the perceptual effect of the interaction between complex tones. Auditory roughness has been suggested by Perakis et al. [49] as an additional metric to describe the sound quality of advanced open rotor aircraft engines, where modulations caused by interaction (beating) between closely spaced complex tones are expected to characterise their noise signature. On the basis of the responses of the participants in the listening tests carried out in this research, the perceived pitch of complex tones (cf. Terhardt's Virtual Pitch Theory [25,50]), and the tonal frequency shift due to the Doppler effect, can help to improve the assessment of perceived effect of the tonal content in aircraft noise. A further improvement of Aures Tonality for assessing the subjective response to tonal content in aircraft noise is to optimise the tonal bandwidth, centre frequency and prominence weighting functions for these particular characteristics.

As mentioned above, there are two exceptions where neither Aures Tonality nor the EPNL Tone Correction play a role in explaining the variance of the PR obtained from the participants' responses which are the GE90-76/85/92B and TRENT1000 engine variants. For the GE90-76/85/92B noise samples, the PR is well correlated to Sharpness, and for the TRENT1000, the PR is well correlated to loudness (i.e. PNL). For these specific aircraft noise samples the participants seemed to focus on the Sharpness and PNL respectively as subjectively dominant features, and the changes in Sharpness and PNL respectively were the main contributors to the variance of the PR. These results are in line with the findings of Torija and Flindell [42] which suggested that the change in the subjectively dominant feature (between different samples) is the most significant contributor to subjective annoyance. Therefore, an interdependent weighting for noise features which takes into account prominence and dominance of the various noise features could aid to assess the subjective response to noise of different aircraft engine types.

Although the participants reported 'high pitch' noise as one of the least preferable characteristics of the aircraft noise samples tested, Sharpness is an important contributor to the PR for only two aircraft engine variants (GE90-76/85/92B and GE90-115B), and does not play any role in explaining the PR of the other four aircraft engine variants. A possible explanation, as described above, is that the tonal content (CFM56-5, LEAP / PW1127G and RB211-5) and PNL (TRENT1000) is the subjectively dominant feature, and therefore the participants did not pay attention to the changes in Sharpness between the specific noise samples of these aircraft engines. These results agree with Gille et al. [48] findings where Sharpness did not correlate with annoyance, even when the participants used the word 'sharp' to describe the aircraft noise samples evaluated. In the same work, Gille et al. [48] found the Total Energy of Tonal Components in the high critical bands x and y ($TETC_{x\text{-}y}$, see Trolle et al. [51] for more information) correlated well with aircraft noise annoyance, and consequently $TETC_{13\text{-}18}$ was

suggested as the metric accounting for the 'sharp' character of the aircraft noise samples evaluated. Further research will investigate the performance of $TETC_{x-y}$ related metrics, or frequency weighting functions assigning more penalty to high frequency tones, for improving the assessment of the perceptual effect of complex tones in aircraft noise.

As shown by Cabell et al. [52], the noise measured for four representative Unmanned Aerial Systems (UAS) was dominated by multiple tones at harmonics of the blade passage frequency (BPF). Christian and Cabell [53] found that EPNL is unable to account for the extra annoyance caused by UAS as compared to road vehicles, and suggested that a more sophisticated tonality method is needed for improving the assessment of UAS noise annoyance. The research presented in this paper, along with the potential improvements to tonal analysis methods described above, can aid in the development of metrics for more accurate assessment of the subjective response to UAS noise.

6. Conclusions

This paper presents the results of a listening test where 35 participants ranked, by order of preference, a series of noise recordings of six variants of aircraft engines representing a contemporary two engine aircraft fleet, with significant differences in the tonal content. The findings of these listening tests suggest that a sophisticated tonality method, such as Aures Tonality, improves on the EPNL Tone Correction in terms of assessing the subjective response to the tonal content of contemporary aircraft noise. In four of the six variants of aircraft engines tested (CFM56-5, LEAP / PW1127G, GE90-115B and RB211-5), the tonal content is the subjectively dominant feature, and Aures Tonality is found to be the main factor explaining the variance of the participants' responses in terms of preference. For the remaining TRENT1000

and GE90-76/85/92B engine variants, the changes in PNL and Sharpness respectively were the main factors explaining the variance of the preference reported by the participants.

A series of limitations of the tonality methods analysed in this paper, and potential improvements for the more accurate assessment of the perceptual effect of complex tonal content in aircraft noise are discussed. The improvements suggested include (i) the optimization of the tonal bandwidth, centre frequency and prominence weighting functions of the Aures Tonality for the particular characteristics of aircraft noise; (ii) a factor to account for the perceptual effect of the interaction between complex tones; (iii) a factor accounting for the perceived pitch of complex tones; and (iv) a factor accounting for the tonal frequency shift due to the Doppler effect. Finally, further work will investigate whether the sound energy of the tonal components in the high frequency region is able to account for the 'sharp' character of the aircraft noise, which have been consistently reported by the participants as one of the least preferable features of the aircraft sounds evaluated; and will investigate the feasibility and performance of an interdependent weighting to account for the subjective dominance of the different character components of aircraft noise.

Acknowledgements

This work was partly supported by the Engineering and Physical Sciences Research Council (grant number EP/M026868/1), and by Innovate UK (Grant No. TSB/113086). All data published in this paper are openly available from the University of Southampton repository at http://doi.org/10.5258/SOTON/D0578.

References

[1] EU Commission - Final Report - Annual Analysis related to the EU Air Transport Market 2016, (2017) https://ec.europa.eu/transport/sites/transport/files/2016_eu_air_transport_industry_analyses_report.pdf

[2] A.J. Torija, R.H. Self, I.H. Flindell. A model for the rapid assessment of the impact of aviation noise near airports. J Acoust Soc Am. 141 (2017) 981-995.

[3] M.N. Postorino, L. Mantecchini. A systematic appoach to assess the ffectiveness of airport noise mitigation strategies. J Air Transp Manag. 50 (2016) 71-82.

[4] P. Gagliardi, L. Fredianelli, D. Simonetti, G. Licitra. ADS-B System as a Useful Tool for Testing and Redrawing Noise Management Strategies at Pisa Airport. Acta Acust United Ac. 103 (2017) 543-551.

[5] A-S. Evrard, L. Bouaoun, P. Champelovier, J. Lambert, B. Laumon. Does exposure to aircraft noise increase the mortality from cardiovascular disease in the population living in the vicinity of airports? Results of an ecological study in France. Noise Health 17 (2015) 328-336.

[6] A. Muzet. Environmental noise, sleep and health. Sleep Med Rev. 11 (2007) 135-142.

[7] W. Babisch, B. Beule, M. Schust, N. Kersten, H. Ising. Traffic noise and risk of myocardial infarction. Epidemiology 16 (2005) 33-40.

[8] J. Dratva, H.C. Phuleria, M. Foraster, J.M. Gaspoz, D. Keidel, N. Kunzli, L.J. Liu, M. Pons, E. Zemp, M.W. Gerbase, C. Schindler. Transportation noise and blood pressure in a population-based sample of adults. Environ Health Perspect. 120 (2012) 50-55.

[9] H.M. Miedema, C.G. Oudshoorn. Annoyance from transportation noise: relationships with exposure metrics DNL and DENL and their confidence intervals. Environ Health Perspect. 109 (2001) 409-416.

[10] K.D. Kryter, K.S. Pearsons. Judges noisiness of a band of random noise containing an audible pure tone. J Acoust Soc Am. 38 (1965) 106-112.

[11] K.S. Pearsons. The effects of duration and background noise level on perceived noisiness. Federal Aviation Agency technical report FAA-ADS-78, 1966.

[12] K.D. Kryter. Concepts of perceived noisiness, their implementation and application. J Acoust Soc Am. 43 (1967) 344-361.

[13] W.C. Sperry. Aircraft noise evaluation. Federal Aviation Agency technical report 550-003-03H, 1968.

[14] A.K. Sahai, M. Snellen, D.G. Simons. Aircraft design optimization for lowering community noise exposure based on annoyance metrics. J Aircraft 54 (2017) 2257-2269.

[15] Federal Aviation Administration (FAA), Noise standards: aircraft type and airworthiness certification, calculation of effective perceived noise level from measured data, in: Federal Aviation Regulations, Part 36, Appendix A2 to Part 36 – Section A36.4, 2002.

[16] K.D. Kryter. The meaning and measurement of perceived noise level. Noise Control 6 (1960) 12-27.

[17] A.K. Sahai, M. Snellen, D.G. Simons. Objective quantification of perceived differences between measured and synthesized aircraft sounds. Aerosp Sci Technol. 72 (2018) 25-35.

[18] R. Angerer, R.A. Erickson, D.A. McCurdy. Development of an annoyance model based upon elementary auditory sensations for steady-state aircraft interior noise containing tonal components. NASA technical report TM 104147, 1991.

[19] S. More. Aircraft noise metrics and characteristics. PhD Thesis Purdue University, 2011.

[20] D. Berckmans, K. Janssens, H. Van der Auweraer, P. Sas, W. Desmet. Model-based synthesis of aircraft noise to quantify human perception of sound quality and annoyance. J Sound Vib. 311 (2008) 1175-1195.

[21] K. White, A.W. Bronkhorst, M. Meeter. Annoyance by transportation noise: The effects of source identity and tonal components. J Acoust Soc Am. 141 (2017) 3137-3144.

[22] S. More, P. Davies. An examination of the influence of tonalness on ratings of aircraft noise. In: Proceedings of the Sound Quality Symposium 2008, Dearborn, Michigan, USA, 2008.

[23] T. Mavris, J. Tai, R. Young, B. Havrilesko. Open rotor noise impact on airport communities. PARTNER Project 35 final report, 2011.

[24] A. Sahai, F. Wefers, S. Pick, E. Stumpf, M. Vorlander, T. Kuhlen. Interactive simulation of aircraft noise in aural and visual virtual environments. Appl Acoust. 101 (2016) 24-38.

[25] E Terhardt, G. Stoll, M. Steewann. Algorithm for extraction of pitch and pitch salience from complex tonal sounds. J Acoust Soc Am. 71 (1982) 679-688.

[26] W. Aures. A model for calculating the sensory euphony of arbitrary sounds. Acustica 59 (1985) 130-141.

[27] ISO 1996-2:2007. Acoustics – Description, measurement and assessment of environmental noise – Part 2: Determination of environmental noise levels, 1987.

[28] T.H. Pedersen, M. Sondergaard, B. Andersen. Objective method for assessing the audibility of tones in Noise Joint Nordic Method – Version 2. AV 1952/99 DELTA Acoustics and Vibration, 2000.

[29] DIN 45681:2005-03, Acoustics – Detection of tonal components of noise and determination of tone adjustment for the assessment of noise immisions, 2005.

[30] A.K. Sahai. Consideration of aircraft annoyance during conceptual aircraft design. PhD Thesis Aachen University, 2016.

[31] Y. Guo, C.L. Nickol, R.H. Thomas. Noise and fuel burn reduction potential of an innovative subsonic transport configuration. In: Proceedings of the 52nd Aerospace Sciences Meeting, National Harbor, USA, 2014.

[32] D.A. McCurdy. Annoyance caused by advanced turboprop aircraft flyover noise. NASA technical report TP-2782, 1988.

[33] S.A. Rizzi, D.L. Palumbo, J. Rathsam, A.W. Christian, M. Rafaelof. Annoyance to noise produced by a distributed electric propulsion high-lift system. In: Proceedings of the 23rd AIAA/CEAS Aeroacoustics Conference, AIAA Aviation Forum, Denver, USA, 2017.

[34] A. McAlpine, P.J.G. Schwaller, M.J. Fisher, B.J. Tester. Buzz-saw noise: Prediction of the rotor-alone pressure field. J Sound Vib. 331 (2012) 4901-4918.

[35] E. Zwicker, H. Fastl. Psychoacoustics: Facts and Model. Springer ed. 1990.

[36] C. Kickson, K. Bolin. Continuous judgment by category-ratio scaling of aircraft noise. Appl Acoust. 84 (2014) 3-8.

[37] G. von Bismarck. Sharpness as an attribute of the timbre of sounds. Acustica 30 (1974) 159-172.

[38] T. Poulsen. Influence of session length on judged annoyance. J Sound Vib 145 (1991) 217-224.

[39] J. Kim, C. Lim, J. Hong, S. Lee. Noise-induced annoyance from transportation noise: short-term responses to a single noise source in a laboratory. J Acoust Soc Am. 127 (2010) 804-814.

[40] A.J. Torija, I.H. Flindell. Listening laboratory study of low height roadside noise barrier performance compared against in-situ field data. Build Environ. 81 (2014) 216-225.

[41] N.J. Versfeld, J.M. Festen, T. Houtgast. Preference judgments of artificial processed and hearing-aid transduced speech. J Acoust Soc Am. 106 (1999) 1566-1578.

[42] A.J. Torija, I.H. Flindell. The subjective effect of low frequency content in road traffic noise. J Acoust Soc Am. 137 (2015) 189-198.

[43] P. Legendre. Species associations: the Kendall coefficient of concordance revisited. J Agr Biol Envir St. 10 (2005) 226-245.

[44] G.W. Corder, DI. Foreman. Nonparametric statistics: A step-by-step approach, Wiley, 2014.

[45] A.K. Sahai, E. Stumpf. Incorporating and minimizing aircraft noise annoyance during conceptual aircraft design. In: Proceedings of the 20th AIAA/CEAS Aeroacoustics Conference, AIAA Aviation Forum, Atlanta, USA, 2014.

[46] A. Minard, C. Lambourg, P. Boussard. Signal-based indicators for predicting the effect of audible tones in the aircraft sound at takeoff. In: Proceedings of INTER-NOISE 2016, Hamburg, Germany, 2016.

[47] B. Barbot, C. Lavandier, P. Cheminee. Perceptual representation of aircraft sounds. Appl Acoust. 69 (2008) 1003-1016.

[48] L.A. Gille, C. Marquis-Favre, R. Weber. Aircraft noise annoyance modeling: Consideration of noise sensitivity and of different annoying acoustical characteristics. App. Acoust. 115 (2017) 139-149.

[49] G.J. Perakis, I.H. Flindell, R.H. Self. Towards roughness as an additional metric for aircraft noise containing multiple tones. Acta Acust United Ac. 99 (2013) 828-835.

[50] E. Terhardt, G. Stoll, M. Seewann. Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions. J Acoust Soc Am. 71 (1982) 671-678.

[51] A. Trolle, C. Marquis-Favre, A. Klein. Short-term annoyance due to tramway noise: Determination of an acoustical indicator of annoyance via multilevel regression analysis. Acta Acust United Ac. 100 (2014) 34-45.

[52] R. Cabell, R. McSwain, F. Grosveld. Measured noise from small unmanned aerial vehicles. In: Proceedings of the NOISE-CON 2016, Providence, USA, 2016.

[53] A.W. Christian, R. Cabell. Initial investigation into the psychoacoustic properties of small unmanned aerial system noise. In: Proceedings of the 23rd AIAA/CEAS Aeroacoustics Conference, AIAA Aviation Forum, Denver, USA, 2017.