Aircraft noise unpleasantness: impact of different signal portions on overall assessment

Antoine Minard
Genesis, France

Bertrand Mellot
Assystem, France

Jean-François Sciabica
Airbus, France

Summary

Reducing aircraft noise in the vicinity of airports is a major concern of Aviation industry. All stakeholders have contributed these past years to important noise level reduction from individual aircraft operations. In order to further improve future aircraft, it appears important to identify and understand the aircraft noise main characteristics responsible for residents' unpleasantness. Numerous studies have used a perceptual approach to assess the impact of some spectral components like Multiple Pure Tones on unpleasantness. Nevertheless, aircraft flyover noise usually includes large variations of sound level and spectral content over its course, with some components only appearing in some signal portions, depending on source directivities. Therefore, it is also important to address the temporal aspects of aircraft flyover noise and their relation to unpleasantness assessment. This paper addresses this question by comparing unpleasantness assessment of different signal portions of aircraft flyover, with various durations. Four reference aircraft sounds were synthesized, two of which contained Multiple Pure Tones. Signal portions were then extracted from these sounds, with different lengths of 5, 10 and 20 seconds, and centered on different parts of the flyovers. The unpleasantness of each created portion was assessed with a comparative evaluation method, where several stimuli are presented at once. Results mainly reveal that none of the 5-second portions is representative of the overall signals in terms of assessed unpleasantness, while only the assessments of the 2 first portions of 10 seconds (beginning and middle parts of the flyover) give unpleasantness ratings comparable to overall assessment. From a practical standpoint, these results offer interesting possibilities for future research on aircraft flyover perception, where shorter flyover excerpts could be considered instead of the overall signal. However, for generalization purposes, these results should be extended to field recordings and a wider variety of aircraft signatures.

PACS no. 43.50.Rq, 43.66.Lj

1. Introduction

Reducing aircraft noise is still a major area for improvement in aviation industry. Noise emission attenuation obtained these past decades through new aircraft technologies has contributed to reduce residents' noise exposure near airports. Nevertheless, the current air traffic growth and the continuous development of residential areas in the vicinity of airports have a significant impact on the noise annoyance perceived by residents. In agreement with the balanced approach to Aircraft Noise Management proposed by the ICAO, and in addition to the development of technologies aiming at reducing the aircraft noise at source or in operation, Airbus is now working on noise annoyance comprehension. Noise annoyance perceived by residents depends not only on acoustical factors (e.g. noise level and frequency content), but also on non-acoustical factors called moderator variables. It was shown that acoustical factors represent 1/3 of the perceived annoyance, whereas the remaining 2/3 are associated with individual sensitivity and social and environmental

factors [1] [2]. Thus, for the aeronautic industry, minimizing the contribution of acoustical factors appears as a necessary step to reduce noise annoyance. This is why understanding, with perceptual studies, the aircraft noise characteristics which are responsible for unpleasantness is required.

The aircraft noise is a complex signal with non-stationary components and a long duration (between 30 and 60 seconds). It is mainly composed of broadband noise (e.g. airframe noise or jet noise), and tonal components (e.g. fan or turbine noise). At take-off, Multiple Pure Tones (also called Buzz Saw Noise) can be heard when the relative speed of the fan at the tip of its blades becomes supersonic. Therefore, these Multiple Pure Tones are responsible for a hearing phenomenon that can be regarded as roughness. Finally, non-stationary effects are created by the trajectory of the aircraft combined with the propagation of its produced noise to the ground (Doppler effect, geometrical spreading, atmospheric absorption, fluctuations associated with atmospheric turbulence, ground reflection and the source directivity of each noise component).

The comprehension of the Multiple Pure Tones impact on global unpleasantness still remains a challenge. Multiple Pure Tones only appear at take-off, at the beginning of the flyover and disappear generally when the aircraft is above the listener. Thier contribution to the global unpleasantness needs to be more deeply understood.
The Sound Quality approach is mainly dedicated to the prediction of the global unpleasantness of a noise source by analyzing perceptual dimensions and identifying indicators to explain them. Sound Quality approach has been already applied on aircraft noise [3]. But this approach does not allow the assessment of the relative influence of separated and/or combined component sources of the global noise.
Continuous assessment methodology has been developed to better understand the perception of non-stationary sounds [4] and seems to be a viable approach to test the impact of Multiple Pure Tones. Recently, it was applied on a set of aircraft noises and compared with global assessment of the unpleasantness [5]. Results show that the maximum instantaneous unpleasantness was mainly correlated with the global unpleasantness. Maximum instantaneous unpleasantness is perceived when the aircraft is above the listener. Therefore, this study has also shown that other portions of the flyover noise have a limited impact on the global unpleasantness, at least for the considered set of sounds. Indeed, in this study, only one of the stimuli had a Multiple Pure Tones component.
The next step would thus be to apply this methodology to a full set of aircraft sounds with Multiple Pure Tones. Nevertheless, this methodology is time-consuming: each listener has to assess the global unpleasantness and to perform the continuous unpleasantness assessment of each noise. As an alternative method, we propose here to compare unpleasantness assessments of different signal portions of aircraft flyovers, with various durations. In section 2 of this paper, the adopted experimental methodology is described. In section 3, results in terms of individual differences, factor analysis and duration comparison are presented. Finally, section 4 draws the main conclusions of this study.

2. Experimental design

2.1. Stimuli

In this study, we are interested in the influence of the duration of stimuli on the assessment of unpleasantness, and which part of the overall flyover signal could best represent the overall evaluations. This is the reason why different shorter extracts were created out of longer reference flyover sounds.
Reference flyover sounds were created with an additive synthesis method, whose principles are explained in [6]. All sounds have in common their broadband noise (encompassing all aircraft sources not considered as “tonal”), whose level was set to match an existing A320 flyover recording, and only differ in their tonal contents (1 or 2 harmonic tones at various frequencies) and their trajectories. Geometric spreading, atmospheric absorption and Doppler effect are recreated. However, neither ground effect nor turbulence are added, in order to limit the impact of other time-related phenomena on sound perception.

Four synthesized sounds served as a basis for creating all the stimuli of the experiment. The first 2 sounds contain only broadband noise and the Blade-Passing Frequency (BPF) component. The other 2 were created by synthesizing a Multiple

Pure Tones (MPT) component and adding it to the first 2 sounds. Finally, one of the 2 initial syntheses was modified by reducing the emergence level of the BPF towards the end of the flyover. These 4 final reference sounds, whose duration was fixed to 20 seconds, are listed in Table I.

Table I. Description of the 4 reference synthesized sounds.

Signal name	**BPF**	**BPF emergence level**	**MPT**
BPF1	1400 Hz	constant	No
BPF1MPT	1400 Hz	constant	Yes
BPF2dec	2800 Hz	decreased towards the end	No
BPF2MPT	2800 Hz	constant	Yes

As examples, Figure 1 shows the spectrograms of the 2 sounds where the MPT component was synthesized (BPF1MPT and BPF2MPT in Table I).

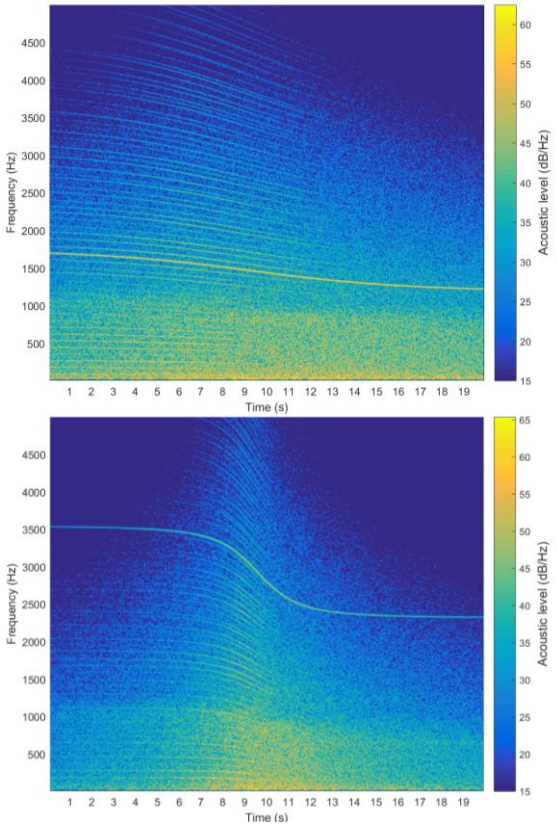

Figure 1. Spectrograms of sounds BPF1MPT (upper panel) and BPF2MPT (lower panel).

From these 4 signals, 10-second and 5-second extracts were considered. For signals of 10 seconds, extracts were taken between:

- 0 s and 10 s;
- 5 s and 15 s;
- 10 s and 20 s.

Twelve signals of 10 seconds are thus considered. For signals of 5 seconds, extracts were taken between:

- 0 s and 5 s;
- 5 s and 10 s;
- 7.5 s and 12.5 s;
- 10 and 15 s;
- 15 and 20 s.

Twenty signals of 5 seconds are thus considered. For the overall 20-second duration, 5 other flyover sounds from the synthesized sound database were added to the 4 reference signals (9 sounds of 20 seconds are thus considered). [1]

2.2. Procedure

The present test used a so-called "comparative evaluation" method, where several sounds are presented and assessed at once (i.e. on the same screen) [7].

However, it was decided that any signal could only be compared to other signals of the same duration. As a consequence, the listening test was composed of 3 "sub-tests", each dedicated to one duration, either 20, 10 or 5 seconds. Furthermore, since the number of sounds of 5 seconds was rather high, they were separated in 2 different sessions performed in sequence by each participant, while making sure that each session included at least 2 sounds with audible MPT (out of 6 possible sounds corresponding to the 3 first signal portions of BPF1MPT and BPF2MPT). Also, common sounds ("anchors") were used in these 2 sessions in order to relate the 2 obtained unpleasantness scales. These anchors were chosen in preliminary listening sessions as predicted least and most unpleasant of the 20 sounds, and gains of -3 dB and 3 dB, respectively, were further applied in order to ensure that they would be assessed as such by the participants.

[1] These additional sounds were mostly considered for comparison with an other non-published study

In the end, each participant underwent 4 comparative evaluation sessions (one with 9 sounds of 20 seconds, one with 12 sounds of 10 seconds, and two with 12 sounds – anchors included – of 5 seconds each), whose order of presentation was randomized each time.

2.3. Participants

Thirty-seven participants (8 women, 29 men, aged 35 on average) volunteered as listeners for this experiment. None of them mentioned any major audition problem.

2.4. Experimental setup

Stimuli were reproduced with Airbus' 3D aircraft noise simulator (see Figure 2). This simulator is composed of 12 loudspeakers, 8 of which are in a circular arc above the participant in order to simulate an overhead trajectory with the VBAP technique (Vector Based Amplitude Panning). A subwoofer was used for low frequencies (10-80Hz) and a representative outdoor background noise was permanently rendered with the 4 other loudspeakers. Each loudspeaker and the subwoofer were individually calibrated at the listening position.

Figure 2. Airbus' 3D aircraft noise simulator

3. Result analysis

This section reports the statistical analysis of the results of the experiment. It mainly consists in observing the individual results as compared to the main tendencies, in order to identify possible outliers, and in studying the effects of the experimental design factors. The output data of the listening test consists in 37 evaluations between 0 and 100 of the unpleasantness of each sound.

3.1. Individual analysis

3.1.1. Anchor evaluations (5-second test)

For the test with 5-second signals, each of the 2 anchors is assessed twice by each participant, because the anchors are presented in both sessions. Almost all participants consistently rated the low anchor at 0 each time and the high anchor at 100 each time (as they were required to rate at least one sound at 0 and one at 100 in each session). Only 4 participants gave different ratings for one of the 4 presentations (2 of each anchor), with deviations between 5 and 10 points from expected ratings.

Because there are so few discrepancies from the expected ratings of anchors, and because these are rather small (at most 10 % of the whole rating scale), it was decided to maintain as they were the 20 ratings (anchors excluded) for each participant, and not to compensate for these peculiarities.

3.1.2. Inter-individual correlations

For each of the 3 considered durations, a Pearson product-moment correlation coefficient was calculated between each pair of individual scales (i.e. participants' scales of answers). A clustering analysis (method UPGMA, see [8] for computational details) was then performed on the obtained correlation matrix. This method allows us to regroup participants into hierarchical clusters according to the similarity of their ratings (in terms of correlation), and makes it possible to potentially identify different rating trends or outliers. The obtained dendrogram, which is the representation of this clustering of the participant panel, is displayed on Figure 3.

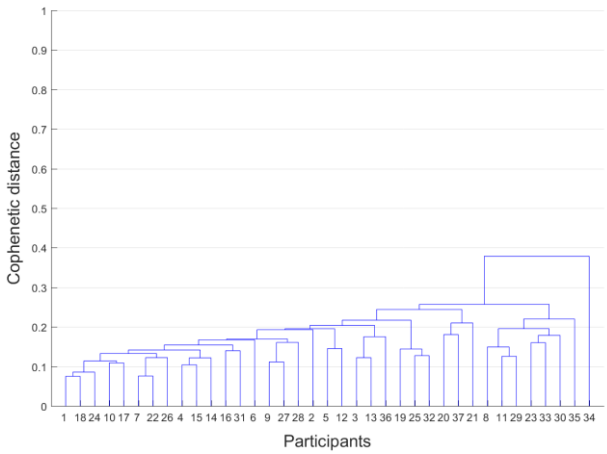

Figure 3. Dendrogram of the participants panel

This figure shows a generally good agreement between participants as identified clusters are positioned quite low in the dendogram. The only exception is participant #34 whose results slightly stand out as being overall less correlated to other participants'. Overall, 92 % of the correlation coefficients are significant (with a type I error rate of 0.05).

To further analyze this divergence of participant #34's results, the same analysis was also applied separately for each of the considered sounds durations. This revealed that participant #34 has always the most or among the most divergent results. These additional analyses also show that general agreement between participants increases for shorter excerpt durations: for 5-s sounds, 96 % of correlation coefficients are significant (with a type I error rate of 0.05), and the mean cophenetic distance (*i.e.* the fusion height of a node linking 2 participants on the dendrogram on Figure 3) is 0.16, while, for 20-s sounds, the percentage drops down to only 30 %, and the mean cophenetic distance is roughly twice as large (see Table 2).

Table 2. Inter-individual agreement metrics.

Sound duration	**Mean correlation coefficient**	**% of significant correlations (α=0.05)**	**Mean cophenetic distance**
20 s	0.44	30 %	0.30
10 s	0.49	48 %	0.26
5 s	0.69	96 %	0.16
overall	0.54	92 %	0.22

3.1.3. Departure from normality

A Jarque-Bera test was applied for each sound (of any duration) to test the data for departure from normality of distribution. The hypothesis of normality is rejected for 9 of the 41 sounds of the test (1 of the 9 sounds of 20 seconds, 2 of the 12 sounds of 10 seconds, and 6 of the 20 sounds of 5 seconds) at the 0.05 type I error rate. However, removing participant #34 from the panel does not reduce this proportion of non-normal distributions (it actually increases to 10 of 41 distributions). For this particular reason, there is no sufficient proof that the results of this participant need to be removed from the data. Thus, all data from the 37 participants were kept in the subsequent analyses.

3.2. Factor analysis

3.2.1. Foreword

The following subsections report the results of analyses of variance. For the sake of clarification, it is important to define the experimental designs considered in these analyses and the factors whose effects on the dependent variable (unpleasantness) are addressed.

The results of the 3 listening tests, corresponding to the 3 durations under consideration (20, 10 and 5 seconds), were addressed separately, because they have distinct experimental designs, which corresponds to different types of ANOVA:

- Sounds of 20 seconds: One-way ANOVA with repeated measures. The only factor is 'Flyover' with 9 levels (9 flyover sounds);
- Sounds of 10 seconds: Two-way ANOVA with repeated measures. The two factors are 'Flyover' with 4 levels (see Table 1) and 'Chunk' with 3 levels (0 to 10 seconds, 5 to 15 seconds, and 10 to 20 seconds);
- Sounds of 5 seconds: Two-way ANOVA with repeated measures. The two factors are 'Flyover' with 4 levels (see Table 1) and 'Chunk' with 5 levels (0 to 5 seconds, 5 to 10 seconds, 7.5 to 12.5 seconds, 10 to 15 seconds, and 15 to 20 seconds).

3.2.2. Verification of assumptions

Normality

Normality was assessed in section 3.1.3, where it was shown that, according to the Jarque-Bera test, 78 % of the overall data (32 of 41 sounds) verifies the hypothesis of normality (with a type I error rate of 0.05).

Homoscedasticity

Homoscedasticity corresponds to the homogeneity of variances. Homoscedasticity was assessed by Levene's test. This test revealed that this hypothesis is violated by the data ($F(19{,}720) = 2.90, p < 0.01$). However, the ANOVA is robust to departure from homoscedasticity so long as sample sizes are similar, which is the case here since repeated measures are considered (by definition, same number of observations – listeners – in each condition).

Sphericity

Finally, sphericity was verified by means of Mauchly's test. For the sounds of 20 seconds the hypothesis of sphericity is verified ($\chi^2(35) = 47.1, p = 0.09$). For the sounds of 10 seconds, it is verified for the 'Flyover' factor ($\chi^2(5) = 8.50, p = 0.13$) and the interaction ($\chi^2(20) = 28.934, p = 0.09$), whereas it is violated for the 'Chunk' factor ($\chi^2(2) = 6.17, p < 0.05$). As for the sounds of 5 seconds, it is only verified for the 'Flyover' factor ($\chi^2(5) = 5.95, p = 0.31$), whereas it is violated for both the 'Chunk' factor ($\chi^2(9) = 41.3, p < 0.01$) and the interaction ($\chi^2(77) = 99.5, p < 0.05$). In case where the hypothesis of sphericity is violated, a Greenhouse-Geisser correction is used in order to compensate for the inflation of the type I error rate.

3.2.3. Analysis of variance

The ANOVA with repeated measures applied on the ratings of the 9 sounds of 20 seconds reveals a significant effect of the flyover on the unpleasantness ($F(8,288) = 31.55, p < 0.001$). This result is not further discussed here, since we are more interested in the comparison of ratings between complete flyovers and those of shorter excerpts, which is the subject of section 3.3.

For the sounds of 10 seconds, the two-way ANOVA with repeated measures reveals significant effects for both the 'Chunk' and the 'Flyover' factors (resp. $F(1.7,62) = 60.48, p < 0.001$ and $F(3,108) = 55.76, p < 0.001$), as well as their interaction ($F(6,216) = 13.22, p < 0.001$). These effects are illustrated in Figure 4, which shows the mean ratings and 95% confidence intervals of the 12 sounds for this duration.

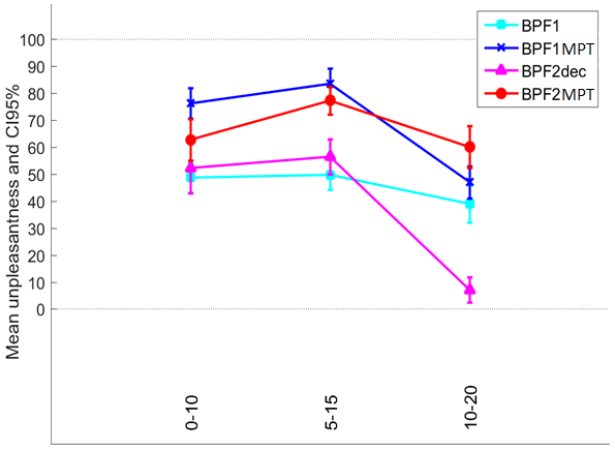

Figure 4. Mean unpleasantness ratings and 95% confidence interval for each of the 12 sounds of 10 seconds

This figure also reveals an effect of the presence of MPT of around 30 scale points for sound BPF1, and 25 scale points for sound BPF2, mostly for the center chunk as the MPT appear later for this sound. The figure also shows that the disappearance of the BPF tone at the end of sound BPF2dec decreases the mean unpleasantness rating of more than 60 scale points.

Finally, for the sounds of 5 seconds the two-way ANOVA with repeated measures reveals significant effects of 'Chunk', 'Flyover' and their interaction (resp. $F(2.4,86.9) = 209.52, p < 0.001$, $F(3,108) = 51.93, p < 0.001$, and $F(7.4,267.9) = 32.23, p < 0.001$).

Figure 5 illustrates these effects. Again, this figure shows both the unpleasantness increase due to the MPT component (maximum increase on the rating scale of roughly 20 points for sound BPF1 and 10-15 points for sound BPF2) and the unpleasantness decrease due to the removal of the BPF tone for sound BPF2dec (around 35 points).

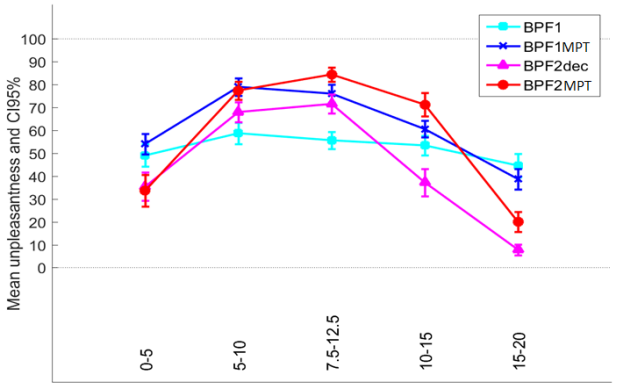

Figure 5. Mean unpleasantness ratings and 95% confidence interval for each of the 20 sounds of 5 seconds

3.3. Duration comparison

This section deals with the question of how similar the ratings are between the different sets of sounds (i.e. different sound durations). The idea behind this comparison is to know whether or not it is possible to obtain reliable ratings (as compared to "long" sounds, i.e. 20 seconds here) with shorter extracts of the sounds. As a consequence, ratings of the 4 reference flyovers, which are present in any of the 3 datasets, will be compared between 20-second sounds and both 5-second and 10-second extracts. Each extract (chunk) is addressed separately. Note however that the sounds of the different durations where not assessed on the same

scale, thus they can only be compared on a relative basis.

For the sounds of 10 seconds, the scatter plots of the overall 20-second ratings and, respectively, the 1st (0 to 10 seconds), 2nd (5 to 15 s) and 3rd (10 to 20 s) chunk ratings, for each of the 4 flyovers, are shown in the 3 graphs on the right of Figure 6.

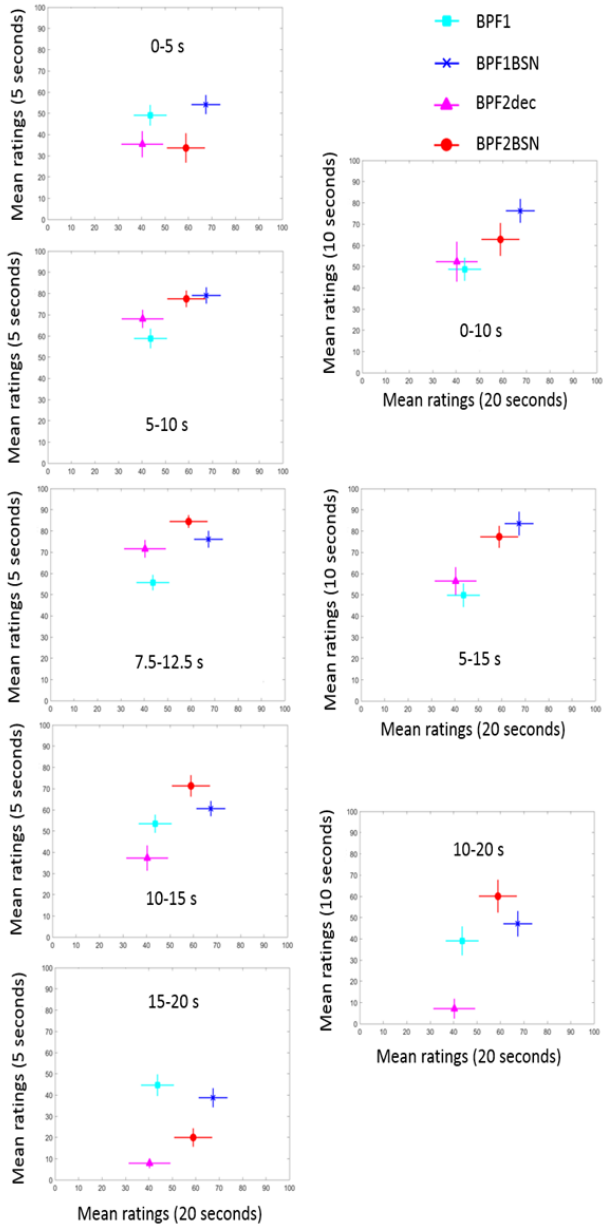

Figure 6: Comparison between mean ratings between 20-second sounds (on the X-axes) and shorter excerpts (on the Y-axes). Graphs on the right allow comparison with 10-second excerpts, and graphs on the left, with 5-second excerpts. In each graphs, horizontal bars represent 95% confidence intervals of mean ratings for 20-second sounds, and vertical bars represent these intervals for shorter excerpts

These graphs indicate that the ratings are quite consistent between the ratings of the 2 first chunks and those of the 20 second sounds. Indeed, the 4 points on 2 first graphs (1st and 2nd chunks) are quite well aligned (the order inversion between flyovers 'BPF1' and 'BPF2dec' seems irrelevant when looking at the confidence interval overlap). On the contrary, the 3rd chunk ratings do not match those of the overall signals.

Using sounds of 10 seconds taken at the beginning or around the center of the flyover would appear to suffice to get reliable results, but this conclusion suffers a low number of points (flyovers). Such a trend should be confirmed with much more examples (at least a dozen), over a larger range of unpleasantness, to be considered as definitive.

For the sounds of 5 seconds, the graphs on the left of Figure 6 show the scatter plots of the overall 20-second ratings and, respectively, the 1st (0 to 5 seconds), 2nd (5 to 10 s), 3rd (7.5 to 12.5 s), 4th (10 to 15 s) and 5th (15 to 20 s) chunk ratings, for each of the 4 flyovers. None of these chunks seems to give comparable results to the overall flyover sounds, since none of these figures shows 4 aligned points. The chunk results were also combined as average ratings of 2 or 3 consecutive chunks, but this did not improve these results.

Using 5-second extracts to assess the unpleasantness of a flyover thus seems inappropriate. The instantaneous unpleasantness seems too dependent on the temporal evolution of the flyover sounds.

4. Conclusion

In this experiment, participants were asked to rate the unpleasantness of 3 sets of flyover sounds of different durations. One of these 3 sets contained 4 reference complete flyovers of 20 seconds. These four sounds were then cut into 3 portions of 10 seconds, and 5 portions of 5 seconds, in order to form the 2 other sets of 12 and 20 sounds respectively.

The analysis of the results mainly revealed that none of the signal portions of 5 seconds was representative of the overall signals in terms of assessed unpleasantness. The 2 first portions of 10 seconds (resp. between 0 and 10 s, and 5 and 15 s of signal) seem to reflect the overall ratings. However, shorter extracts give higher agreement rates between participants, which means that ratings averaged across participants are more reliable. From a methodological standpoint, these results could be highly beneficial, as the rather long duration of stimuli makes perceptual experiments on aircraft flyover sounds practically

difficult. Nonetheless, the apparent reliability of 10-second extracts observed here lacks consistency over a larger set of flyovers and should be confirmed further before considering the use of signals of 10 seconds as a reliable replacement for usual flyover signals of at least 20 seconds.

Additionally, this study shows strong rating differences when noise components such as the MPT are varied. These differences transpire for any duration of signal, but are not always comparable between long and short signals. More specifically, 5-second extracts taken at the center of the signal, where the sound level is maximum, show diverging trends from 20-second signals (center graph on the left of Figure 6). This appears as a contradiction with the results obtained in [5], where the maximum instantaneous unpleasantness ratings seemed to suffice to explain overall assessments. Two possible explanations (not necessarily exclusive) can be hypothesized:

- The divergence observed here is only due to the use of sounds where the MPT are synthesized, whereas this component was hardly considered in [5].
- The divergence comes from the fact that participants, when assessing a given 5-second extract, have no immediate knowledge of what the sound was before that extract. This would mean that the assessment of a flyover sound produced at each instant is not independent of assessments at past instants.

Further research work is needed to explore these 2 possibilities. A first step towards a better understanding of the relation between the temporal evolution of flyover sounds and unpleasantness would be to reproduce the experiment conducted in [5] on sounds with more variety, and specifically sounds containing tones or MPT.

References

[1] R. Guski, U. Felscher-Such, R. Schuemer: The concept of noise annoyance, how international expert see it. Journal of Sound and Vibration, 223(4). (1999).

[2] T. J. Schultz: Synthesis of social surveys on noise annoyance. J. Acoust. Soc. of Am., 64 (1978)

[3] Eu-Project SEFA (Sound Engineering For Aircraft) Publishable Summary. Available here in March 2018 https://trimis.ec.europa.eu/project/sound-engineering-aircraft#tab-docs (2007).

[4] P. Susini, S. McAdams: Psychophysical validation of proprioceptive device by cross-modal matching of loudness. Acta Acustica united with Acustica, 86, 515-525 (2000).

[5] G. Lemaitre, J.-F. Sciabica, S. Moal, L. Vion, M. Zekri, S. Hourcade, P. Boussard : Développement d'un dispositif d'évaluation continue du désagrément acoustique généré par un survol d'avion. Proceedings of Congrès Français d'Acoustique 2014. (2014)

[6] A. Minard, S. Hourcade, C. Lambourg P. Boussard, Controllable sound simulations of aircraft flyovers, Proc. Internoise 2015, San Francisco, USA, 2015.

[7] P. Chevret, E. Parizet. An efficient alternative to the paired comparison method for the subjective evaluation of a large set of sounds. Proc. 19th ICA, Madrid, Spain, 2007.

[8] P. Legendre, L. Legendre: "Numerical Ecology". Elsevier, Amsterdam, 1975.