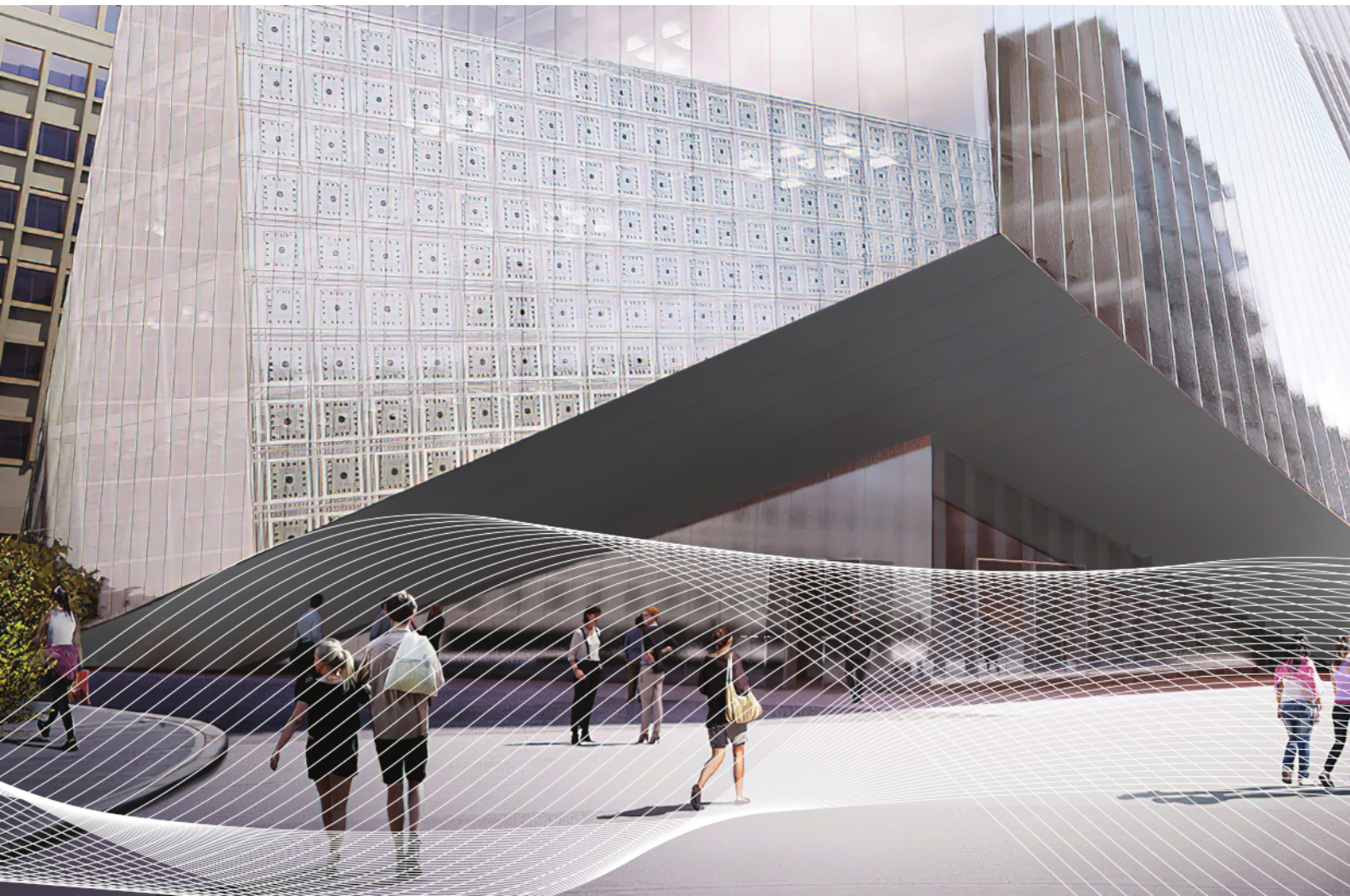
Perceptual evaluation of clarinet sounds in the context of instrument making

Ossen El Sawaf, Tom Colinot, Sabine Meunier, Jacques Chatron, Michael Jousserand, Philippe Guillemain

17e Congrès Français d'Acoustique
27-30 avril 2025, Paris

Perceptual evaluation of clarinet sounds in the context of instrument making

O. El Sawaf [a], T. Colinot [b] S. Meunier [a] J. Chatron [a] M. Jousserand [b] et P. Guillemain [a]
[a] CNRS - LMA, 4 impasse Nikola Tesla, CS 40006, 13453 MArseille, France
[b] Buffet Crampon, 5 rue Maurice Berteaux, 78711 Mantes-La-Ville, France



This study describes the construction and results of a listening test in which clarinet sounds were evaluated according to several qualities represented by adjectives. Six adjectives were selected as the most shared among several interviews with expert musicians and makers. The sounds were recorded specifically for the study and selected, cut into short excerpts, and presented to the participants. Only certain sounds can be set apart as exhibiting consensus on certain adjectives. One sound in particular is evaluated very homogeneously, which seems to be linked to its frequency content. A PCA showed redundancies between adjectives. Of the six adjectives chosen, *bright* and *clear*, as well as *warm* and *round*, are very close. It seems that sounds can be assessed on two scales: bright - warm and nasal - sweet. The search for an objective indicator for these adjectives, using the MIR toolbox, was unsuccessful.

1 Introduction

In the context of instrument making, professional testers and musicians use specific terms to describe the sound of the instrument. The link between these terms and objective descriptors of the sound signal are rarely well established, with the exception of brightness, which appears to be related to spectral centroid ([1]). This study attempts to establish links between signal descriptors and adjectives related to instrument sound quality. This presentation describes the construction and results of a listening test set up to evaluate different pre-recorded clarinet sounds according to several adjectives (bright, warm, clear, sweet, nasal, round). We will present the construction of the adjective corpus, the choice of parameters used, the construction of the test and the results obtained.

2 Choosing adjectives based on experts' language

First, we wanted to determine which adjectives are most commonly used by experts. For this purpose, interviews were conducted with 14 French experts: clarinet and oboe professional musicians, clarinet testers; clarinet, oboe and saxophone technicians. They were asked to freely explain the qualities that in their opinion are important and make an instrument a well-crafted one (free verbalization such as in [2, 3]. During the interviews, they were encouraged to elaborate on the terms they used and to give as precise definitions as possible. All interviews were recorded, transcribed and analyzed in order to extract words related to various categories such as sound produced, ease of sound production or ease of play. Each definition given by the experts was studied to determine which words were most precise and which ones meant the same thing.

From this analyses, six words emerged as the most shared between the experts. Work by [4] suggests that even if certain words are not the most common among people, there may be a more general consensus on their definition. However, the aim here is not to propose definitions of words, and the audience we are addressing is made up entirely of experts. In this respect, it seems appropriate to let the experts express themselves and to collect the most widely shared words among them. The selected words are presented in table 1 in French with their English translations provided.

It should be noted that this project started right before the COVID19 pandemic. As such, it was very complicated to go and meet Buffet Crampon experts which were not allowed to play in front of people (playing was to be done in a closed room by themselves). At the time of the experiment, covid regulations were still going, which influenced the experiment. In this situation, it was decided to carry out an experiment using headphones, even though the initial idea was also to evaluate words that designated playing qualities. However, it would then have been necessary to have the participants play the instruments, which was not possible. One consequence of this is that the terms selected are only those related to the sound characteristics of the instrument.

French	Brillant	Chaud	Clair	Doux	Nasillard	Rond
English	Bright	Warm	Clear	Sweet	Nasal	Round

Table 1: Adjectives selected for the listening test

3 Recording and selecting stimuli

3.1 Recording

Clarinet recordings of five different clarinet models, played by three Buffet Crampon experts were made. Each of the experts was asked to play a descending and ascending chromatic scales tied and untied at 100 BPM, an excerpts from Mozart's clarinet concerto covering almost the entire clarinet ambitus at 57 BPM and a free choice piece.

Recording were made in two ways : a record was made using an ORTF stereo with the microphones (2 x NEUMANN KM 184) placed right above the expert's head and another recording with two microphones placed in front of the horn (DPA 4099) and halfway above the clarinet's body (AKG C480 + CK61 capsule). Recordings were made with Adobe Audition using a RME Babyface interface and a RME Octamic preamp. The total recording time was 39 minutes and 18 seconds.

3.2 Processing and segmenting

Taken separately, the microphones placed at the horn and above the body do not pick up the full spectrum of the instrument, so these two signals were mixed together, with the mix validated by an experienced sound engineer. After careful listening, it was these mixes that were chosen for the experiment, as the recording from the ORTF microphones sounded less authentic because of their location.

Stimulus number	Duration	Notes
1	0.6	1 note, drawn from the scale (A)
2	0.5	1 note, drawn from the scale (D)
3	0.6	2 notes, drawn from Mozart's concerto (E)
4	2.3	3 notes, drawn from Mozart's concerto (B)
5	0.6	1 note, drawn from the scale (B)
6	0.5	1 note, drawn from the scale (C)
7	0.9	3 notes, drawn from Mozart's concerto (B)
8	0.4	1 note, drawn from the scale (C)
9	0.3	1 note, drawn from the scale (A)
10	0.4	1 note, drawn from the scale (C)
11	0.5	2 notes, drawn from Mozart's concerto (D)
12	0.5	3 notes, drawn from Mozart's concerto (D)
13	0.7	3 notes, drawn from Mozart's concerto (E)
14	0.4	2 notes, drawn from Mozart's concerto (E)
15	1.1	2 notes, drawn from Mozart's concerto (C)
16	1.7	3 notes, drawn from Mozart's concerto (C)

Table 2: Stimuli description, sorted in ascending order of pitch. For each stimulus, its duration and the recording from which it was taken are indicated. The letter in brackets indicates which clarinet is used in the choosen recording.

The recordings were meticulously studied in order to isolate the cleanest passages, i.e. free of interfering noise such as the sound of the keys or the musician's blowing, for example. From the passages retained, groups of 1 to 3 notes were isolated in order to construct a set of stimuli covering the range of the clarinet and made up of passages with different expressive qualities.

A number of points were taken into account when constructing the set of stimuli: 1) the stimuli were to be short enough to make the duration of the test acceptable for professionals with little time available, 2) they had to cover the full range of the clarinet and 3) it had to be ensured that there were differences in timbre between the excerpts. Three sets of 16 sounds excerpts were proposed and a preliminary listening test between lab members was made to determine which stimuli set was the most suitable for the experiment. The final set was ultimately made up of 16 sounds from the three original sets, each signal consisting of a couple of notes with duration between 0.3 to 2.3s. Although we are aware that this results in very short signals, this choice was made as a compromise between having enough stimuli to cover the whole range of the clarinet and keeping the listening test to around 30 minutes (taking into account several listenings of each signal). The description of the stimuli is given in table 2.

4 Listening and evaluation test methodology

As the experiment was initially going to be performed by Buffet Crampon's expert, duration was a very important parameter and should not be too long. Thus, the use of a method where each sound would have been evaluated one by one for each term retained or a paired comparison for each term were rejected. The work of [5] compared six different methods and showed that a mixed method where all stimuli

Figure 1: Interface for the training part of the test

are available for evaluation at the same time for one given adjective was almost as precise as a paired comparison for a third of the duration. This method allows participants to evaluate each sound and also to compare them with each other.

The test consisted of evaluating the 16 sounds according to one adjective at a time, with the 6 adjectives presented in random order. For each adjective, the 16 sounds were represented by a button that could be clicked to play the associated sound. Next to each button was a continuous scale allowing the sound to be evaluated by placing a cursor between 'not at all [X]' and 'very [X]' (X being the adjective currently evaluated). To the right of the scales, there were also two boxes ('yes' and 'no') for the participant to indicate whether the sound heard was relevant to the adjective proposed. The idea was that if a signal was always declared 'irrelevant' for a given adjective, the physical characteristics that could be derived from this signal would not, a priori, play a part at all in the qualitative aspect linked to this word.

Figure 1 shows a screenshot of the interface for the training part of the test. The training part only differs in the number of stimuli (3 instead of 16) and words (3 instead of 6). At the top of the interface the currently evaluated term is shown. In figure 1 the word is "Riche" (translated to "Rich" in English), this adjective isn't in the table 1 as it was only used in the training part of the test. The buttons on the left ('SON 1', 'SON 2' and 'SON 3') play the randomly associated stimulus when clicked. Then, the listener can evaluate the sound they heard by placing a cursor along the continuous scale. All sounds could be played as many times as wanted and the cursor could also be adjusted as many time as wanted by the listeners. When the listener had finished evaluating the sounds, a "Changement d'adjectif" button (which translates to "Adjective Cahnge") allowed them to switch to a new adjective chosen at random from the list, with each "SOUND" button corresponding to a new, randomly ordered sound. The button for changing to a new adjective was only active if all the stimuli had been played at least once.

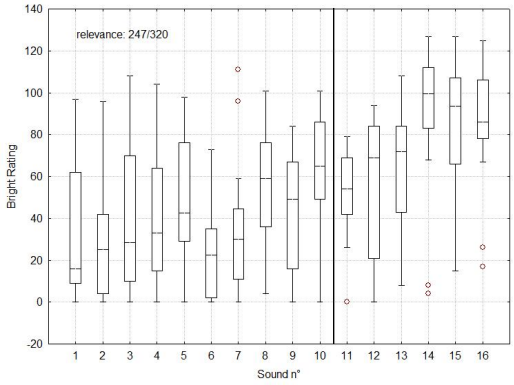


Figure 2: Boxplot representation of *bright* ratings. Relevance indicate how many evaluations were noted as relevant. The bold vertical line separates first and second register. The first and third quartile are limiting the box, and the median is presented as a bar within it. Whiskers include all non-outlier values that do not fall into the box ; circles are outliers (between 1.5 and 3 times the interquartile value away from the first or third quartiles), and asterisks extreme values (more than 3 times the interquartile value away from the first or third quartiles).

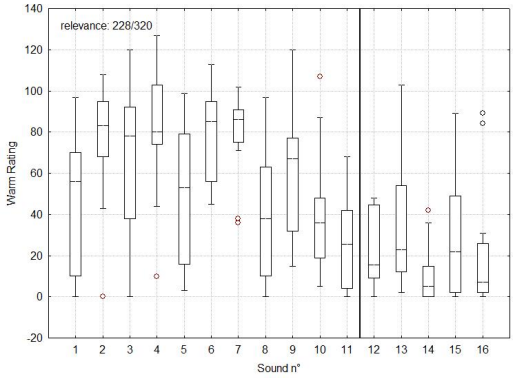


Figure 3: same as figure 2, for adjective *warm*.

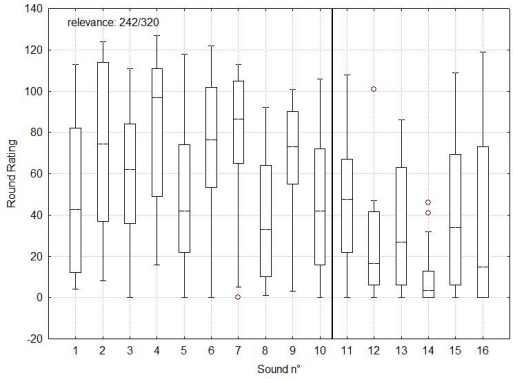


Figure 4: same as figure 2, for adjective *round*.

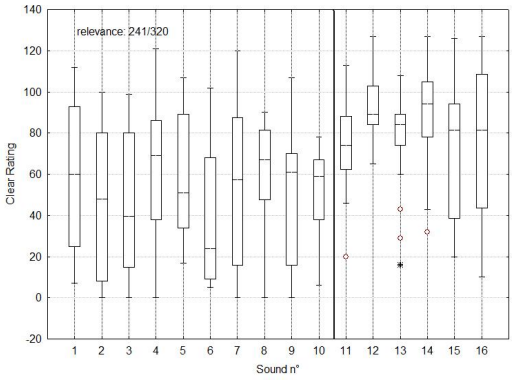


Figure 5: same as figure 2, for adjective *clear*.

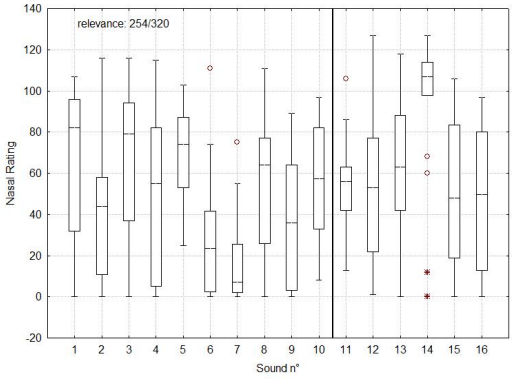


Figure 6: same as figure 2, for adjective *nasal*.

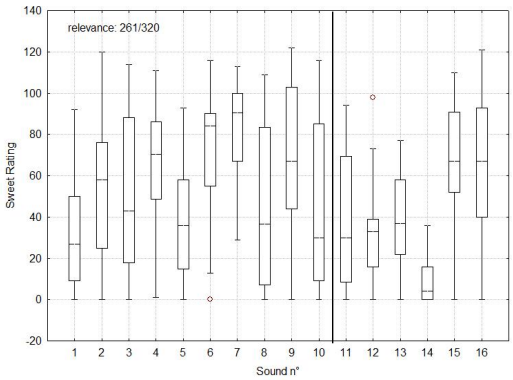


Figure 7: same as figure 2, the adjective *sweet*.

5 Results

Twenty people took part in the experiment and only 3 did not play an instrument. Among the musicians, 1 was a professional musician, 6 semi-professionals and there were 5 clarinetists (1 professional, 2 semi-professional and 2 amateurs).

A numerical value was assigned to each evaluation, the scale being coded on 128 values from 0 ("not at all [X]") to 127 ("very [X]"). All results include only sound/adjective evaluations that were marked as *relevant* by participant.

Evaluations marked as *irrelevant* suffer a strong bias towards lower values, because when marking the evaluation as *irrelevant* participants often left the cursor untouched (which reads as 0 in final results).

5.1 Dependance to pitch

As a first step in the analysis, the stimuli are analyzed according to their average pitch. This allows to split them into two categories, based on which acoustic mode of the clarinet resonator sustains the oscillation. The first category regroups sounds of the first register, i.e. the lowest 19 notes of the clarinet where the heard note corresponds to the first mode of the resonator (below written B♭4). The second category regroups any note higher than that, i.e. the higher registers whose fundamental frequency correspond to the second or third mode of the resonator. Figures 2 to 7 displays boxplots of all evaluations with stimuli sorted by average pitch, and split into the first register and higher registers.

The *bright* (figure 2) rating globally increases with pitch, without a clear demarcation or change in trend between the first register and the higher registers.

The *warm* rating (figure 3) is almost always very low for higher registers. As an indication of this, one can notice that the highest median in the higher register is lower than the lowest first register median. A similar but less marked tendency appears for evaluated *round* (figure 4). However, for both adjectives, there is no clear monotonous trend inside of each register.

The *clear* rating (figure 5) is almost always very high for higher register. This is illustrated by the lowest higher register median being higher than the highest first register median.

The *nasal* and *sweet* ratings (figures 6 and 7 respectively) don't show any specific tendency with regard to registrer, they don't seem to be dependant on pitch.

The fact that there is no monotonous relationship with the pitch leads us to say that it is indeed not the pitch that is analysed by the participants to differentiate the stimuli but, a priori, the timbre.

5.2 On consensus among subjects

Before looking at the question of consensus among subjects, we want to determine whether each adjective discriminated the stimuli. In other words, whether the stimuli were rated significantly differently for each adjective. As we are not sure whether the scales were used in the same way by the different participants, we perform Friedman ANOVAs for this purpose, which analyzes the ranks and not directly the ratings. Friedman's ANOVA showed a significant effect of stimulus on *bright*, *warm*, *clear*, *sweet* and *round* (Chi2 (9, 15) = 80.16, p <0.0001; Chi2 (7, 15) = 52.4 , p <0.0001; Chi2 (7, 15) = 35. 9, p =0.0018; Chi2 (10, 15) =60.05 , p <0.0001; Chi2 (9, 15) =55.64 , p <0.0001), but showed no significant effect of stimulus on *nasal* (Chi2 (6, 15) = 23.97, p <0.06). It should be noted that missing values (when the adjective was indicated as not relevant for the stimulus) were not taken into account, and thus the participant excluded from the analysis. These results lead us to believe that the stimuli were adapted to elicit differences in sensation in the participants for almost all the adjectives.

As shown in the boxplots in figures 2 to 7, most stimulus/adjectives combinaison exhibit a very large distribution of values. However, many of the combinations display consensus such as :

- stimulus #16 for adjective bright (figure 2)
- stimulus #7 for adjectives sweet (figure 7) and warm (figure 3)
- stimulus #11 for adjective nasal (figure 6)
- stimulus #12 for adjectives clear (figure 5) and round (figure 4)

All these exemples of combinations show a relatively short box indicating some degree of consensus between participants.

Stimulus #14 is probably the most interesting of the study. Indeed, for all the adjectives, it is always this stimulus whose median is the highest or the lowest, and this with relatively short whiskers each time. This indicates a strong consensus regarding its evaluation, clarinetists and non-clarinetists included.

Many comments can be made on a single adjective accumulating such a large number of interesting statistical features. First, this point to a semantic overlap of the chosen adjectives: it is possible that our six adjectives hold some redundancy when applied to clarinet sounds. If we focus on the illustrative objective of the study, i.e. the search for sounds that help define or illustrate given adjectives, the present results are a strong incentive to choose stimuli very carefully.

5.3 Correlation with objective signal descriptors

As mentioned above, our results lead us to believe that some adjectives might evoke overlapping definitions in our participants. Therefore, a principal component analysis was carried out, the results of which are shown in the figure 8.

The PCA shows that two factors explain most of the variance (factor 1 : 73,37%, factor 2 : 18,54%). In particular, it shows that adjectives *Bright* and *Clear* are very similar, as well as *Round* and *Warm*.

Correlations between objective descriptors from the MIR toolbox ([6]) and four adjectives were then computed (following the PCA results, *bright* and *warm* were used to represent also *clear* and *round* respectively). The *brightness* descriptor, defined as the proportion of spectral energy above 1500 Hz, correlates positively with ratings for *bright*, *clear* and *nasal*, and negatively with ratings for *warm*, *sweet* and *round*. Correlation plots with associated *r* values are shown in figures 9 to 12. Although the r-values obtained are fairly good, the fact that all the adjectives are correlated with the same descriptor (*Brightness*) tends to indicate that this descriptor cannot provide an objective assessment of the perceived quality of the stimuli used since it does not distinguish between the different adjectives.

It is interesting to note that stimulus #14, which always obtained an extreme rating, is also the highest in terms of the *brightness* descriptor. This could be taken as indicating that the proportion of energy in a particular part of the spectrum

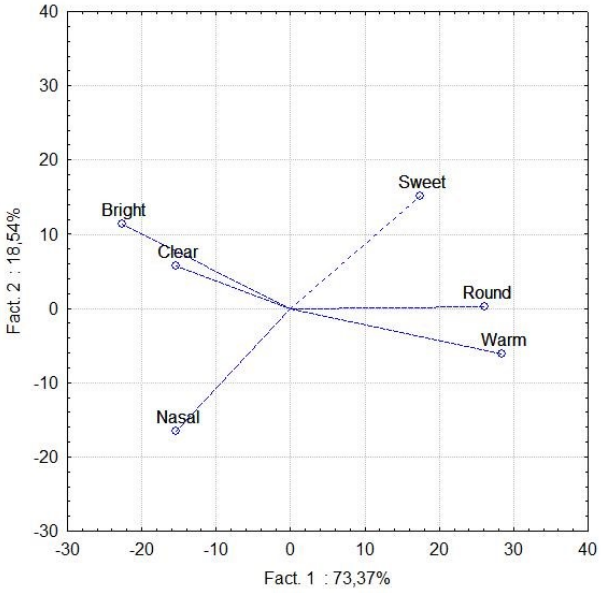


Figure 8: Result for the PCA.

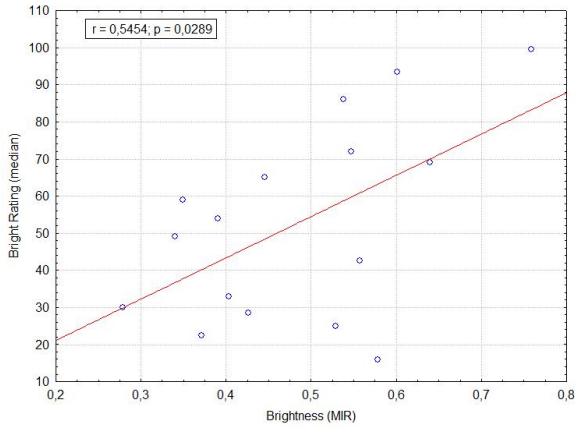


Figure 9: Correlation between *bright* rating and the objective descriptor *brightness* of the MIR toolbox.

clearly plays a part in the evaluation of the signal. However, focusing simply on the proportion of energy above a given frequency seems to override all other cues, and fails to distinguish between different adjectives.

Other descriptors from the MIR toolbox were computed and compared to the perceptual evaluations. For instance, *Spectral centroid* also holds the information for high-frequency energy, and was somewhat correlated to *Nasal* (r = 0.45). However, oustide of *Brightness*, no descriptors exhibited clear tendencies.

6 Discussion and conclusion

Stimulus duration was the subject of a lot of oral comments by participants during the tests, especially by expert clarinet players. They mentioned finding it hard to form a judgment on sound quality with such short stimuli. They reported lacking musical context, even going so far as saying that what they heard "did not sound like a clarinet".

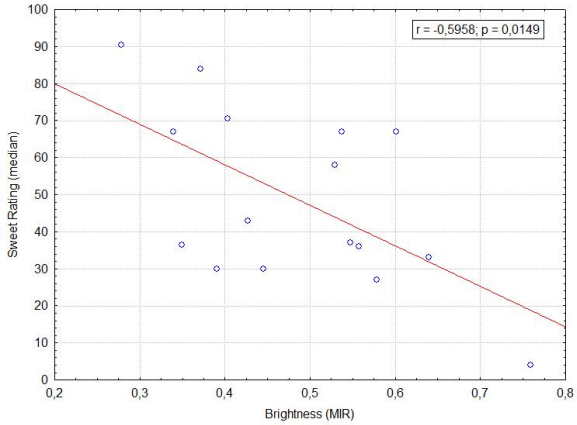


Figure 10: Correlation between *sweet* rating and the objective descriptor *brightness* of the MIR toolbox.

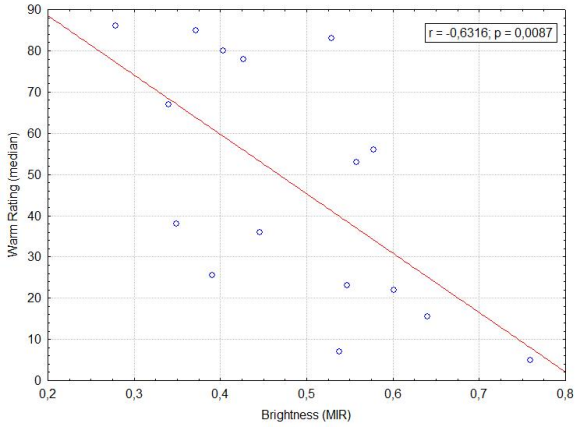


Figure 11: Correlation between *warm* rating and the objective descriptor *brightness* of the MIR toolbox.

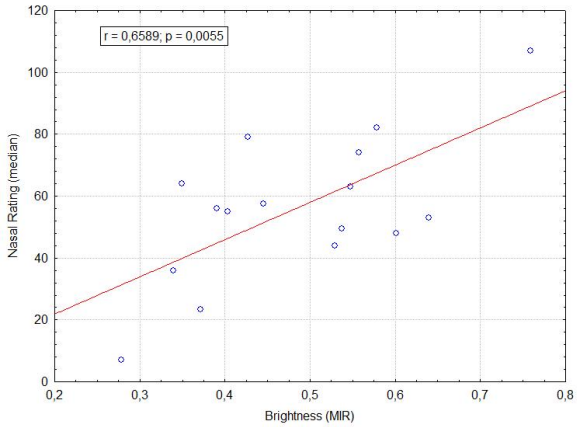


Figure 12: Correlation between *nasal* rating and the objective descriptor *brightness* of the MIR toolbox.

Such comments were less frequent from non-clarinetists. Overall, when asking musicians about their main instrument, providing musical context seems necessary for them to form a solid opinion on the perceived quality of the sounds. Of course, this requires longer stimuli and therefore a longer test procedure, which could be more tiring for participants. This duration issue, as well as the fact that a single stimuli was evaluated very strongly for all adjective and by both clarinetists and non-clarinetists further highlights the importance of stimulus selection.

The PCA result show that the evoked qualities of *warm* and *round* seem to overlapped, as well as those for *bright* and *clear*. However, it also shows that two dimensions seems to explain differences between adjectives and that *sweet* and *nasal* exists as opposite on the same dimensions. The same observation can be made for *bright* and *round*. This may indicates that a couple of objectives descriptors might be enough to explain the perceived qualities of our stimuli.

While several sound/adjectives combinations show agreement between participants, many show blurry results instead. This may come not only from a major difference of experience but also because when a musician talks about their instrument, they also consider the haptic aspect, particularly for a wind instrument like the clarinet where the musician is physically coupled to the instrument. The authors believes that a lot of words they used may hold meaning regarding sound as well as haptics. For that reason, the experiment's original design involved listening *and* playing tests as a way to encompass all these aspects. However, as mentionned, the project happened during the COVID19 pandemic which had a huge impact : for one thing, we couldn't envisage a test where instruments would be shared. The results of the present study, however, further underline the importance of a test in which participants would play the instrument under study.

References

[1] S. McAdams, Perspectives on the contribution of timbre to musical structure, *Computer Music Journal* **23(3)**, 85-102 (1999).

[2] M. Garnier, N. Heinrich, D. Dubois, M. Castellengo, J. Poitevineau and D. Sotiropoulos, Etude de la qualité vocale dans le chant lyrique, *Scolia* **20**, 151-169 (2005).

[3] P. Cheminée, C. Gherghinoiu and C. Besnainou, Analyse des verbalisations libres sur le son du piano versus analyses acoustiques, *Colloque interdisciplinaire de musicologie* (2005).

[4] A. Faure, Des sons aux mots, comment parle t-on du timbre musical ?, *Thèse de doctorat*, Paris, EHESS (2000).

[5] E. Parizet, N. Hamzaoui and G. Sabatie, Comparison of Some Listening Test Methods: A Case Study, *Acta Acustica united with Acustica* **91**, 356-364 (2005).

[6] O. Lartillot and P. Toivinainen, A matlab toolbox for musical feature extraction from audio, *International conference on digital audio* **237**, 244 (2007).

[7] P. Susini, O. Houix, N. Misdariis, P. Faramaz and G. Pellerin, The speak project: A collaborative platform for presentig and sharing sound lexicons, *16e Congrès Français d'Acoustique* (2022).