STUDY OF THE PERCEIVED QUALITY OF SAXOPHONE REEDS BY A PANEL OF MUSICIANS

Jean-François Petiot
Pierric Kersaudy
LUNAM Université, Ecole Centrale de Nantes
CIRMMT, Schulich School of Music, McGill University
Petiot@irccyn.ec-nantes.fr
Pierric.Kersaudy@eleves.ec-nantes.fr

Gary Scavone
Stephen McAdams
CIRMMT, Schulich School of Music, McGill University
gary@music.mcgill.ca
smc@music.mcgill.ca

ABSTRACT

The subjective quality of cane reeds used on saxophones or clarinets may be very different from one reed to another even though the reeds have the same shape and strength. The aim of this work is to study the differences in the subjective quality of reeds, assessed by a panel of musicians. The work focuses mainly on the agreement of the panel of musicians, the reliability of the evaluations and the discrimination power of the panel. A subjective study, involving 10 skilled musicians, was conducted on a set of 20 reeds of the same strength. Three descriptors were assessed: *Brightness*, *Softness*, and *Global quality*. The ratings of the musicians were analyzed using sensory data analysis methods to estimate the agreement between them and the main consensual differences between the reeds.

Results show that for *Softness* and *Brightness,* the agreement between the musicians is important and that significant differences between the reeds can be observed. For *Global quality*, the inter-individual differences are more important. The performance of the panel in providing reliable assessments opens the potential for an objectification of the perceived quality.

1. INTRODUCTION

For a saxophone player, the quality of a reed (a piece of cane that the player places against the mouthpiece) is fundamental and has big consequences on the quality of the sound produced by the instrument. The experience of saxophone players roughly shows that in a box of reeds, 30% are of good quality, 40% are of medium quality and 30% are of bad quality. The only indicator a musician can see on a box of reeds is the strength, which is usually measured by the maker by submitting a static force on a particular location from the tip. The reeds are then classified according to the strength measured. But this strength is not representative of the perceived quality of the reed. According to musicians, there are many differences among the reeds in a given box. But it is still difficult to understand which physical or chemical properties govern the perceived quality. The control of reed quality remains an important problem for reeds makers, because of the important variability of this natural material (arundo donax) and of the huge number of influencing factors. A thorough study of the perceived quality of reeds, and more generally of musical instruments, necessitates two categories of measurements on a set of products: subjective assessments (given by musicians or listeners) [1] and objective measurements (chemical or physical), made on a set of instruments [2]. The principle is next to uncover (with statistical methods) a model for predicting subjective dimensions from the objective measurements.

In [3], optical measurements were used to assess the vibrational modes of clarinet reeds, which had been correlated with the quality of the reeds as judged by musicians. The authors suggested different patterns of vibrations that should be representative of good reeds.

In [4], B. Gazengel and J.P. Dalmont proposed two categories of physical measurements to explain the behavior of a tenor saxophone reed (in vivo during playing, and in vitro with a testing bench measuring the mechanical frequency response). Additional studies using these measurements showed that the perceived strength of a reed can be explained by the estimated threshold pressure in the musician's mouth, and that the perceived brightness correlates with the high-frequency content of the sounds [5, 6]. But these results were based on a small set of reeds (12) and used only one musician to assess their quality. They were limited to simple correlations between subjective variables and objective measurements and need to be confirmed.

The main difficulty in the study of the perceived quality of musical instruments is to get subjective assessments from musicians that are reliable and representative enough of the subtle interaction between the musician and the instrument. Many uncontrolled factors may influence this complex interaction. The subjective ratings of a "subject" may be non-reproducible, context-dependent, semantically ambiguous, and dependant on cultural and training aspects of the musician. To get representative data, it is necessary to find an acceptable trade-off between realistic playing conditions and artificial assessments of stimuli that could be oversimplified and then too caricatural. And to trust the data, it is necessary to control the assessments with repetitions and with several independent assessors. In this context, experimental protocols and data analysis techniques developed in sensory analysis can be very useful [7]. A number of statistical analysis methods are proposed to assess the evaluations of subjects and the panel's performance in descriptive analysis tasks [8].

In a previous paper [9], we defined a predictive model of tenor saxophone reed quality with PLS regression.

This model was based on a set of 20 reeds and a panel of 10 musicians.

This paper is the continuation of that work. It is centered particularly on the study of the performances of the panel of musicians. We propose to evaluate the inter-individual differences and to assess the reliability of the subjective assessments.

The paper is organized as follows: Section 2 presents the details of the experiment carried out with a set of reeds and a panel of musicians for the subjective study. Section 3 is dedicated to the presentation of the results of the subjective study. The agreement between the different assessments is presented. The last section presents the general conclusions and discusses the contribution of this study.

2. MATERIAL AND METHOD

2.1 Reed samples

The set of 20 reeds for tenor saxophone all had the same cut, strength and brand (Classic Vandoren, Strength 2.5). There was no preliminary selection of the reeds; they all came from 4 commercial boxes of 5 reeds each. The objective here is to estimate the perceived differences in 20 “similar” reeds.

Ten musicians participated in the subjective tests. They were all skilled saxophonists (students or professionals, with more than 10 years of practice). For the sake of consistency, all subjects used the same mouthpiece during the study (Vandoren V16 T7 Ebonite), however they were asked to play on their own tenor saxophone. These subjective tests took place at CIRMMT (Center for Interdisciplinary Research in Music Media and Technology) in Montreal, Quebec, Canada in May 2012.

2.2 Subjective evaluation of the reeds

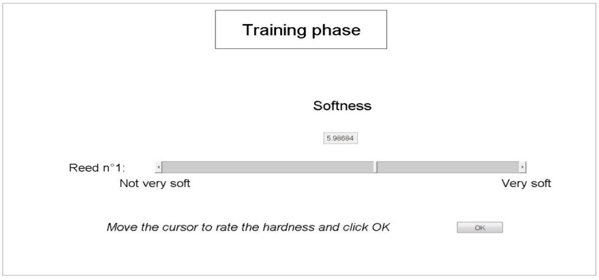
In subjective tests, different semantic dimensions are generally defined to assess the differences between products [10]. For saxophone reeds, interviews of saxophonists have shown that the most frequent dimensions relate to “ease of emission”, “quality of sound”, or “homogeneity”. We proposed three subjective descriptors to assess the reeds:

- The *Brightness* of the sound produced with the reed,
- The *Softness* of the reed, which corresponds to the ease of producing a sound,
- The *Global quality* of the reed.

The test was divided into 3 phases: a training phase, an evaluation phase, and the filling out of a questionnaire concerning the mouthpiece, reed, saxophone and musical style the musicians usually play, as well as their past experience.

The training phase was proposed to help the subjects understand the meaning of the two descriptors *Softness* and *Brightness* and to verify their use of the scale. “Anchor reeds”, located at the extremes of the *Softness* scale, were proposed, and recorded sounds with different *brightnesses* were proposed. The method is inspired from the training phase described in [11]. Finally, subjects were asked to rate 3 quite different reeds on the interface, to train them in the use of the scales and to verify their discrimination.

The evaluation phase used a graphical interface to assess the reeds. The musician was asked to play each reed and to assess each descriptor on an unstructured continuous scale (example in figure 1).



Training phase

Softness

5.98684

Reed n°1:

Move the cursor to rate the hardness and click OK

Figure 1. Continuous scale for the assessment of *Softness*

The reeds were presented to the subject in an order following a Williams Latin square in order to control the order and carry over effects. Given that we have 20 reeds and 10 subjects, the presentation plan was perfectly balanced. The assessments were repeated two times in two independent blocks. For each of the 10 subjects, the subjective data consists of 2 arrays of quantitative values (one per repetition). The arrays have 20 rows (one per reed) and 3 columns (one per descriptor).

The sensory panel consisted of J=10 assessors who judged I=20 products during K=2 sessions using M=3 attributes. The assessment of product *i* by assessor *j* during session *k* according to descriptor *m* is denoted Y_{ijk}^{m}.

3. RESULTS AND DISCUSSION

3.1 Individual performances of the assessors

This section focuses on the individual performances of the assessors, to whether the results of some subjects should be discarded. We use in this section the principles of the GRAPES method [12], which has been developed to assess the performances of a panel of experts in sensory analysis. It provides graphical representations of assessors’ performances. We will focus on the different uses of the scale, the reliability of the subjects, their repeatability and their discrimination capacity.

3.1.1 Use of the scale

Two quantities can be computed to compare the use of scales by assessors. LOCATION$_j$ is the average of the scores given by assessor *j* (equation 1); SPAN$_j$ is the average standard deviation of a score given by assessor *j* within a session (equation 2). It represents the average magnitude used by the assessor to discriminate the products.

$$LOCATION_j = Y_{.j.} \tag{1}$$

$$SPAN_j = \frac{1}{K}\sum_k \left[\frac{\sum_i (Y_{ijk} - Y_{.jk})^2}{(I-1)}\right]^{1/2} \tag{2}$$

N.B. We use a synthetic notation for the representation of the mean: considering the evaluation Y_{ijk} (see section

2.2), the notation $Y_{.j.}$ means the mean of evaluations Y_{ijk} over the indices i (product) and k (session).

Figure 2 presents SPAN$_j$ vs LOCATION$_j$ for the different descriptors for subjects S1 to S10.

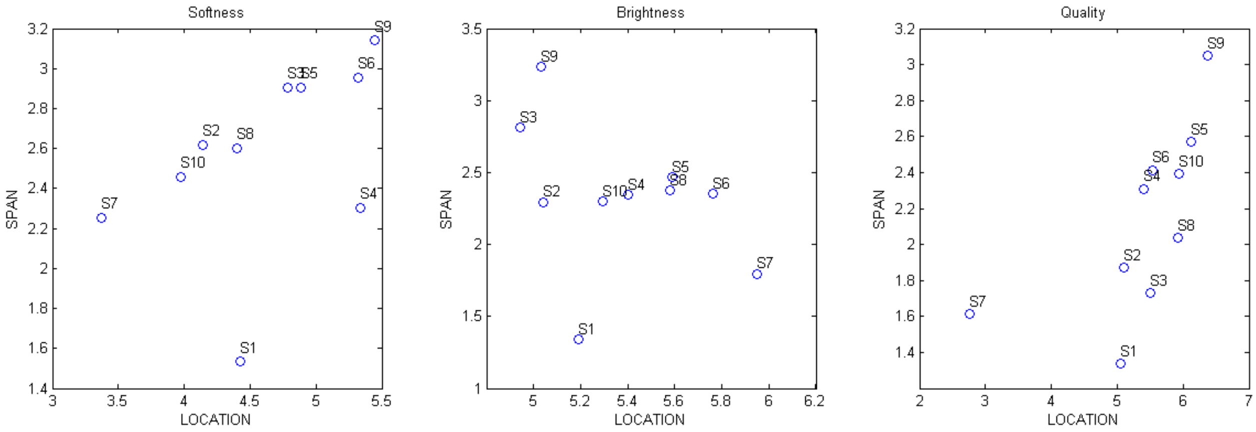


Figure 2. Plot of SPAN$_j$ vs LOCATION$_j$ for each subject and each descriptor.

The results show that subject S1 uses a small range for all the assessment (the SPAN is very small) and subject S7 globally dislikes all the reeds and assesses them as not soft (LOCATION is low for this subject).

3.1.2 Reliability of the subjects and influence of the session

Two coefficients can be computed to assess the performance of each subject for each descriptor concerning their reliability and the influence of the different repetitions.

The unreliability ratio, labeled UNRELIABILITY$_j$, represents the measurement error of the subject, relative to the average magnitude used for the ratings. It is given by equation (3):

$$UNRELIABILITY_j = \frac{\left[\frac{1}{(I-1)(K-1)}\sum_{i,k}\left(Y_{ijk}-Y_{ij.}-Y_{.jk}+Y_{.j.}\right)^2\right]^{1/2}}{SPAN_j} \qquad (3)$$

The DRIFT_MOOD$_j$ (equation 4) is the between-sessions error relative to the average magnitude used for the ratings (expressed in SPAN units). It represents the deviation of the ratings of the subject across the sessions.

$$DRIFT\_MOOD_j = \frac{\left[\frac{1}{K-1}\sum_{k}\left(Y_{.jk}-Y_{.j.}\right)^2\right]^{1/2}}{SPAN_j} \qquad (4)$$

Figure 3 represents, for each descriptor, the performance of the subjects according to DRIFT_MOOD and UNRELIABILITY.

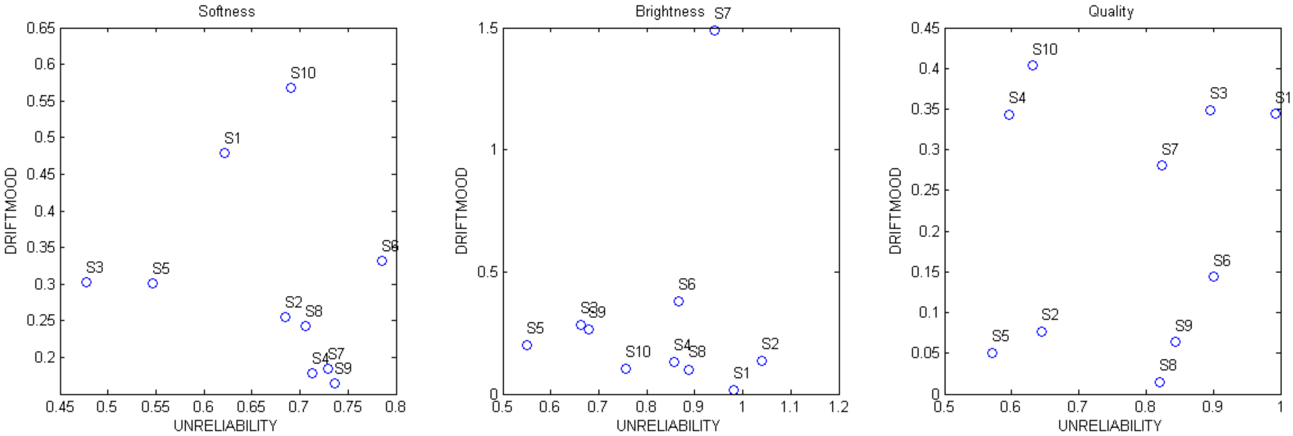


Figure 3. Plot of DRIFT_MOOD$_j$ vs UNRELIABILITY$_j$ for each subject and each descriptor.

For *Softness*, S6 is the least reliable and S3 and S5 are the most reliable. S10 deviates the most between the 2 sessions (high DRIFT_MOOD). For *Brightness*, S2 is the least reliable and S5 is the most reliable. S7 deviates the most between the 2 sessions. For *Quality*, S1 is the least reliable and S5 is the most reliable.

We can conclude that S5 is a particularly reliable subject. We can also see that the worst value of unreliability for *Softness* is lower than most of the values for *Brightness*.

This means that most subjects (S6, S4, S8, S1, S2, S7) are less reliable for *Brightness* than for *Softness*. This result is in accordance with the feedback of the subjects during the tests, who indicated having more difficulty assessing *Brightness* than *Softness*.

These graphs are interesting to verify the quality of the individual assessments in order to detect possible unreliability or misunderstanding in the ratings. In our panel, no subject is particularly identified as unreliable in the assessment.

3.2 Global performance of the panel

3.2.1 Agreement between the assessors

The agreement between the assessors in their evaluation of the reeds can be estimated by consonance analysis, a method based on a principal component analysis (PCA) of the assessments. A description of this method can be found in [13]. To study the agreement for each descriptor (independent of the sessions), the repetitions are merged vertically (repetitions are considered as different products). A standardized PCA is performed on the matrix $Y^m(2\,I\,\mathrm{x}\,J)$ (equation 5):

$$Y^m = \begin{bmatrix} Y_1^m \\ Y_2^m \end{bmatrix} \tag{5}$$

A perfectly consensual panel would consist of assessors who rate the reeds in the same way. In this case, the first component of PCA would account for a very large variance. The more the panel is consensual, the more the arrows of the assessors point in the same direction. The percentage of the variance explained by the first principal component is considered as an indicator of the consonance of the panel. The results of the PCA of the matrices Y^m are given in figure **4** for each descriptor. In this PCA, the variables are the assessors (S1 to S10) and the individuals are the reeds.

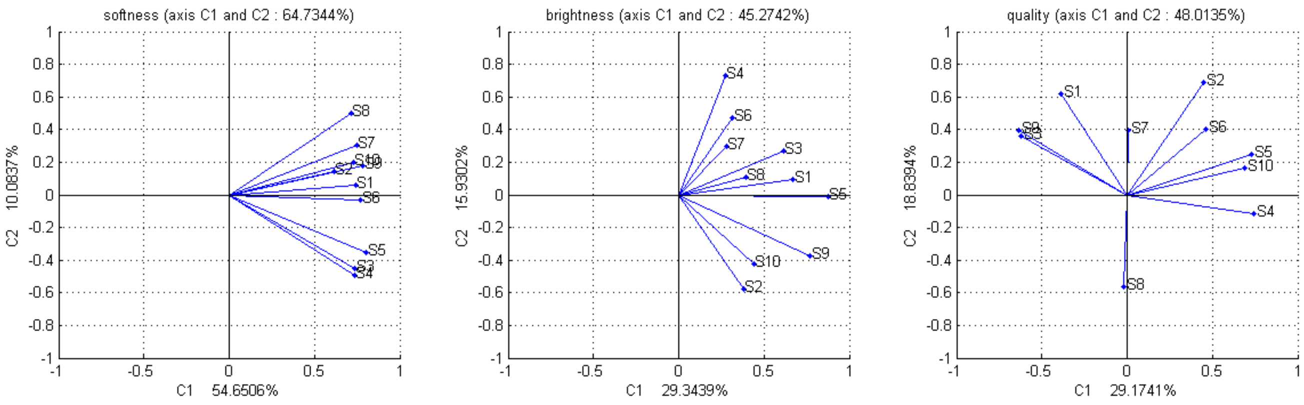


Figure 4. Consonance analysis for each descriptor: plot of the first two factors of the PCA (plane of the variables)

To evaluate more precisely the strength of the consensus for each descriptor, we can use indicators such as the Consonance C defined by equation 6 [13]:

$$C = \frac{\lambda_1}{\sum_{r=2}^{J} \lambda_r} \tag{6}$$

where J is the components number in the PCA (here the number of assessors), and λ_r is the rth eigenvalue of the covariance matrix associated with the rth component in the PCA. So this indicator emphasizes the weight of the first principal component and considers the higher dimensions as error or noise. It can be compared to a signal/noise ratio. We can also use the percentage of the total variance explained by the first principal component as an indicator to estimate the consonance of the panel.

The consonance ratio C and the variance accounted for by the first factor are given in Table **1**.

Descriptor	Consonance C	% Variance first PC
Softness	1.2	54.6%
Brightness	0.4	29.3%
Global quality	0.4	29.2%

Table 1. Results of consonance analysis for the panel of subjects.

The highest agreement is obtained for the descriptor *Softness*. The opinions of the assessors are convergent and the agreement is strong. For *Brightness*, the agreement is weaker, even though no assessor is very discordant.

For *Quality*, the agreement is the weakest. This is rather normal, given that *quality* is strongly related to the preference of the saxophonist, and that the tastes of the musician can be very diverse. Subjects S1, S3, and S9 are rather opposite to the rest of the panel; subject S8 is independent of the general trend according to preference. Given this result, we will have to analyze the *global quality* separately from the two other descriptors and for different groups of subjects.

3.2.2 Discrimination power of the panel

A general method to estimate the discrimination power and reproducibility of a panel of assessors is the Analysis of Variance (ANOVA). It is used in sensory analysis to study the differences between products and, more generally, to test the statistical significance of qualitative factors [14].

The assessment of the product *i* by assessor *j* during session *k* is denoted Y_{ijk} (*i*=1 to I, number of products, *j*=1 to J, number of assessors, *k*=1 to 2, number of sessions). A model for the whole panel (equation 7) is proposed, taking into account the reed effect α_i, the session effect γ_k, and the reed*session interaction $\alpha\gamma_{ik}$:

$$Y_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + \epsilon_{ijk} \qquad (7)$$

In this model, we don't introduce the subject effect because we consider that we don't have enough degrees of freedom to estimate correctly the contribution of the subject effect, the reed effect, the session effect and the associated interactions in the same model. As a matter of fact, the reed effect determines the discriminant power of the panel, and the reed*session interaction determines the repeatability of the panel. Consequently, the subject becomes a random variable in the model and gives us more analysis power. An ANOVA model is fit for each descriptor. The results of the ANOVA for the whole panel are given in Table **2**.

	p-value		
Source of variation	*Softness*	*Brightness*	*Quality*
Reed	<0.001	<0.001	0.028
Session	<0.001	0.005	0.34 (n.s.)
Reed*Session	0.21 (n.s.)	0.88 (n.s.)	0.96 (n.s.)

Table 2. Results of ANOVAs for the three descriptors (p-value)

The reed effect is significant for all the descriptors (p <0.05), which signifies that the panel discriminated the reeds well. The reed*session interaction is not significant for all the descriptors (p >0.05), which means that there is no significant disagreement in the panel from one session to another. The session effect is significant for *Softness* and *Brightness*. It is a sign of a slight change in the use of the scale between the two sessions. Given that the reed effect is significant, we consider that the panel of assessors is discriminant/repeatable enough to aggregate the data in a consensual evaluation, representative of the reeds.

3.3 Subjective characterization of the reeds

3.3.1 Descriptive analysis

The mean value and the standard deviation of the assessments have been computed for each descriptor. The mean values are represented in figure **5** for *Brightness* and figure 6 for *Softness*.

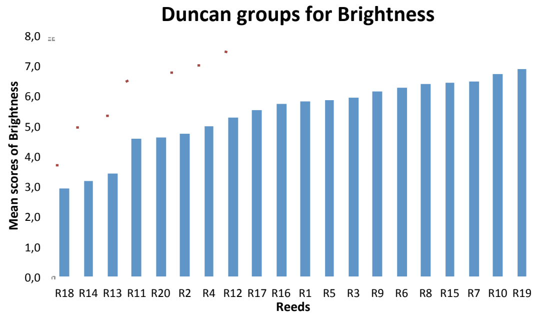


Figure 5. Mean value of *Brightness* and Duncan groups (multiple comparison test – p = 5%)

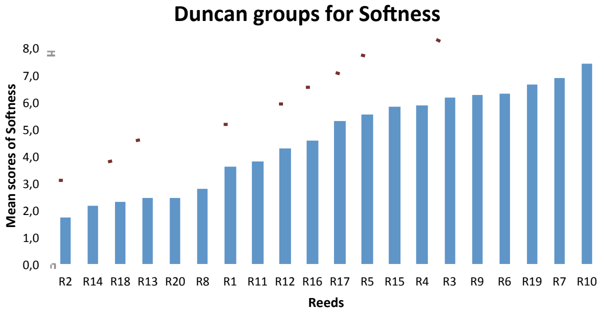


Figure 6. Mean value of *Softness* and Duncan groups (multiple comparison test – p = 5%)

Significant differences between the reeds are evaluated by a Duncan multiple comparison test. Depending on the attributes, the Duncan multiple comparison test enables discrimination between 7 (*Brightness*) and 9 (*Softness*) non-overlapping groups of reeds. The Duncan groups (5% level) are represented by the pieces standing under the same horizontal. Figures 5 and 6 detail the differences between reeds that are significant for each attribute. The test confirms that the discrimination between the reeds is better for *Softness* than for *Brightness*.

The average position of the reeds (R1 to R20) is given in Figure **7**.

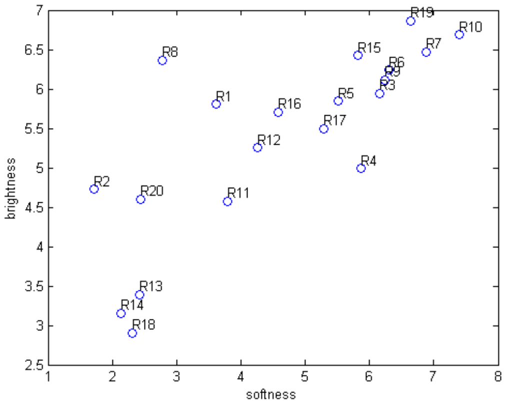


Figure 7. Position of the reeds according to *Softness* and *Brightness* (average configuration)

R10, R7, R19 are the most soft and bright reeds; R14, R18, R13 are the least soft and bright reeds. There is also a correlation between the two descriptors *Brightness* and *Softness*: a bright reed is also generally soft.

3.3.2 Analysis of the global quality

We showed in section 3.2 that the agreement between the assessors for the attribute *Quality* was relatively weak, and that discordant subjects should be considered. For these reasons, the subjects were partitioned according to quality. Let us consider the assessments of quality in the matrix Y^3 of dimension (2I×J), which considers the repetition as additional variables (variable = reed*session).

A cluster analysis with Hierarchical Ascendant Classification has been made on the matrix Y^3. We performed the cluster analysis on the row data (not centered nor reduced) because we consider that the verbal anchoring of the scale gives a meaning to the scores and the mean. The distance used for the HAC is the Euclidian distance and the linkage rule is the Ward criterion (variance criterion). The dendrogram of the classification is presented in figure 8 (grouping of the subjects).

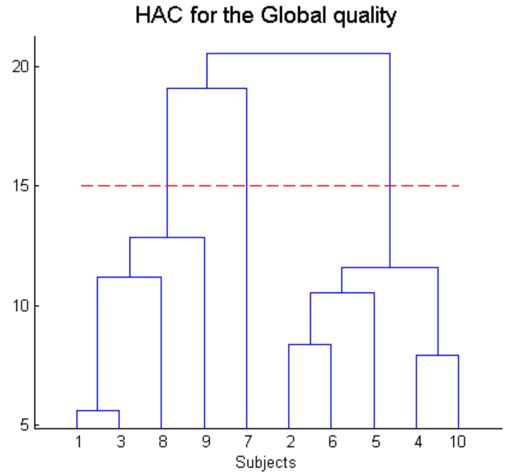


Figure 8: Dendogram of the HAC according to the global quality ratings for the mean of the 2 sessions

3 clusters can be formed:

- Group1: S1 S3 S8 S9.
- Group2: S2 S6 S5 S4 S10.
- Group3: S7.

The average scores of reed quality for the two main groups 1 and 2 are given in figure 9.

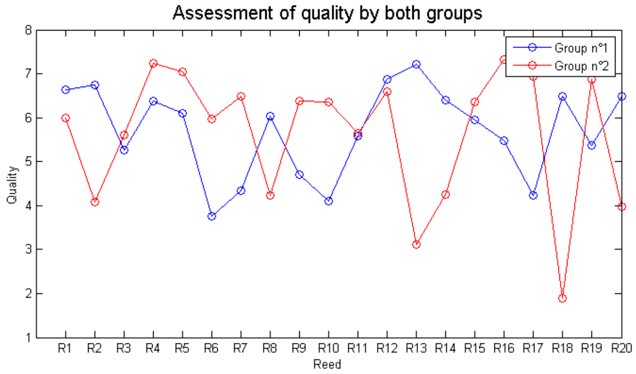


Figure 9: quality scores for the 2 different groups

Group 1 and 2 have mainly conflicting opinions on reeds R13 and R18 (most segmenting reeds). Group 1 (typical subject S3) appreciates R13 and R18, whereas Group 2 (typical subject S10) dislikes them.

We tried to characterize both groups with external information concerning the subjects, obtained from the questionnaires, but no feature of the musicians seems to clearly characterize the groups. However it seems that most of the musicians in group 1 play hard reeds and most of the musicians in group 2 play soft reeds. But we can't generalize this because of the small number of musicians we had. This seems logical, because the biggest differences we can see between the two groups are on the softest reeds or on the hardest reeds. For example we can see big differences for the reeds R2, R13 and R18, which are perceived as the hardest reeds, and we also see big differences for the reeds R10 and R17, which are perceived as soft reeds.

4. CONCLUSIONS

This paper presented an analysis of the subjective assessments of a set of 20 saxophone reeds. Three descriptors were assessed by a panel of 10 musicians: *Softness*, *Brightness* and *Global Quality*.

The results show that the agreement between the subjects is more important for *Softness* than for *Brightness*. For these two descriptors, with the proposed task, the musicians were able to provide discriminant assessments and significant differences between the reeds are observed.

Differences between the musicians concerning the perceived quality necessitated the definition of subgroups of musicians. These differences are normal and due to the differences in personal tastes of the musician.

Future work will consist in using machine learning technique to model the subjective assessments by objective measurements.

Acknowledgments

The authors would like to thank the 10 musicians of the Schulich School of Music, McGill University, Montreal for their participation in the subjective tests, as well as Bruno Gazengel from LAUM for his advices.

5. REFERENCES

[1] Pratt R.L., Bowsher J.M. "The subjective assessment of trombone quality". Journal of Sound and Vibration 57, 425-435 (1978).

[2] Pratt R.L., Bowsher J.M. "The objective assessment of trombone quality". Journal of Sound and Vibration **65**, 521-547 (1979).

[3] F. Pinard, B. Laine, and H. Vach. Musical quality assessment of clarinet reeds using optical holography. *The Journal of the Acoustical Society of America*, 113:1736, 2003.

[4] B. Gazengel and J. Dalmont, "Mechanical response characterization of saxophone reeds," in proceedings of Forum Acusticum, Aalborg, June-July 2011.

[5] B. Gazengel, J.-F. Petiot and E. Brasseur, "Vers la définition d'indicateurs de qualité d'anches de saxophone," in proceedings of 10ème Congrès Français d'Acoustique, Lyon, April 2010

[6] B. Gazengel, J.-F. Petiot and M. Soltes, "Objective and subjective characterization of saxophone reeds," in *proceedings of Acoustics 2012*, Nantes, april 2012.

[7] Marjorie C. King, John Hall, and Margaret A. Cliff. A comparison of methods for evaluating the performance

of a trained sensory panel. Journal of Sensory Studies, 16(6):567 582, 2001.

[8] Zacharov N., Lorho G. What are the requirements of a listening panel for evaluating spatial audio quality? Spatial audio & sensory evaluation techniques, Guildford, UK, 2006, April 6-7.

[9] Petiot J-F., Kersaudy P., Scavone G., McAdams S., Gazengel B. Modeling of the subjective quality of saxophone reeds. Proceedings of ICA 2013, June 2013, Montreal, Quebec, CANADA.

[10] A. Nykänen, O. Johansson, J. Lundberg, J. Berg. Modelling Perceptual Dimensions of Saxophone Sounds. Acta Acustica united with Acustica, Volume 95, Number 3, May/June 2009 , pp. 539-549 (11).

[11] S. Droit-Volet, W. Meck and T. Penney, "Sensory modality and time perception in children and adults," Behavioural Processes, vol. 74, pp. 244-250, 2007.

[12] P. Schlich, "GRAPES: A method and a SAS program for graphical representation of assessor performances," Journal of sensory studies, vol. 9, pp. 157-169, 1994.

[13] G. Dijksterhuis, "Assessing Panel Consonance," Food Quality and Preference, vol. 6, pp. 7-14, 1995.

[14] T. Couronne. “A study of Assesors’ Performance Using Graphical Methods”. Food, Quality and Preference Vol. 8, No. 5/6, pp. 359-365, 1997.