Ecole Centrale de Nantes　　*Université de Nantes*　　*Ecole des Mines de Nantes*

MASTER AUTOMATIQUE ET SYSTEMES DE PRODUCTION

SPECIALITE ARSI

Année 2010 / 2011

Thèse de Master

Présentée et soutenue par :

PIERRIC KERSAUDY

le 03/10/2012

à l'Institut de Recherche en Communications et Cybernétique de Nantes

TITRE

Toward a predictive model of the subjective quality of saxophone reeds

JURY

Président :	Jean-François Lafay	Professeur, IRCCyN, Nantes
Examinateurs :	Saïd Moussaoui	Maître de Conférences, IRCCyN, Nantes
	Marie-Françoise Lucas	Maître de Conférences, IRCCyN, Nantes
	Eric Lecarpentier	Maître de Conférences, IRCCyN, Nantes
	Sébastien Bourguignon	Maître de Conférences, IRCCyN, Nantes
Directeur(s) de thèse :	Jean-François Petiot	Professeur, IRCCyN, Nantes

Laboratoire : Institut de Recherche en Communications et Cybernétique de Nantes – UMR CNRS 6597

Acknowledgements

I would like to thank McGill University for greeting me in the MusicTech Lab. I would like to thank Jean François Petiot who gave me this opportunity to work with him on this interesting problematic. I could not do all this work without his help. Thanks to Steve McAdams, Bruno Gazengel and Gary Scavone for their advice and their help on this project. That was a pleasure to work among the MusicTech lab and what I learned in Montréal will be useful for my future. And finally, thanks to the 10 subjects for participating in the subjective study I led.

Table of contents

1 Introduction

For a saxophone player, the quality of a reed is fundamental and has big consequences on the quality of the sound produced by the instrument. The reed is a piece of cane that the player place against the mouthpiece. And when the player blows, the reed auto-oscillates and a sound is produced. The experience of the saxophone players shows that in a box of reeds, 30% are of good quality, 40% are of mean quality and 30% are of bad quality.

Usually, the only indicator a musician can see on a box of reeds is the strength. The strength is usually measured by submitting a static force on a particular location from the tip. The reeds are then classified according to the strength measured. But the strength is not representative of the quality of the reed. Even for the stiffness of the reed (which should be linked to the strength), there are many differences between the reeds into a box. So the strength is not able to explain the differences between the reeds into a box.

In [1], B. Gazengel and J.P. Dalmont proposed two methods to determined objective variables that describe the behavior of a tenor saxophone reed. On the one hand, they performed “in vitro” measurements using a mechanical bench, and on the other hand, they performed “in vivo” measurements by measuring acoustic pressure of the saxophone and the pressure in player mouth. Several studies have also been led on this topic by B. Gazengel and J.F. Petiot about the correlation between some subjective descriptors and some objective variables[2][3]. A high correlation has been found between the perceived strength of the reed and the objective variable called Pressure Threshold coming from “in vivo” measurements. These studies have the disadvantage to taking account of the assessment of only one subject for the subjective part, so the subjective assessment can be easily questioned. The problem is that there are few subjects (only one), so the data coming from this study are not reliable enough. Then, the correlation task has been stopped at the simple correlation study which doesn’t allow predicting the subjective features of the reeds by the objective ones.

This work proposes to deal with the question using a subjective/objective approach (Figure 1). The principle is to lead on a given set of reeds, a subjective study and an objective study on the other hand. After that we try to correlate the results of both studies, and more generally to set up a machine learning approach to built a model explaining the subjective ratings by the objective measurements.

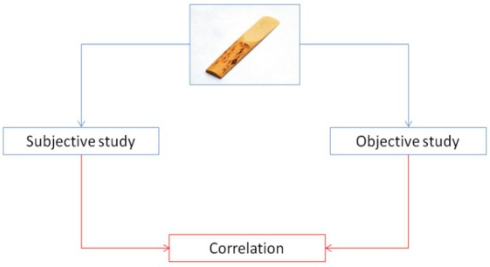


Figure 1: Diagram of the main steps of a subjective/objective study.

In this work, we propose to lead a thorough subjective study with several assessors in order to be confident in the results of this subjective part. Then we propose a method to deal with the variability of the objective measurements. Finally, we propose to go further than the simple correlation and to build a model that predicts the subjective features of the reeds using several objective variables.

2 Material and methods

2.1 Description of the subjective tests

To study the correlation between the perceived quality and the objective measurements, we first need to assess the subjective features of the reeds. We begin with the following product space: 20 reeds for tenor saxophone of the same cut, the same strength and the same 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 was to verify the perceived differences within given box. 10 subjects participated in the subjective tests. They were all skilled saxophonists (students or professionals). To guarantee consistency, all subjects used the same mouthpiece during the study (a Vandoren V16 T7 Ebonite). However, they were asked to play on their own tenor saxophone. Three subjective descriptors were assessed:

- The brightness of the sound produced with the reed.
- The softness of the reed, which corresponded to the ease of producing a sound.
- The global perceived quality of the reed.

This test was divided into 3 phases.

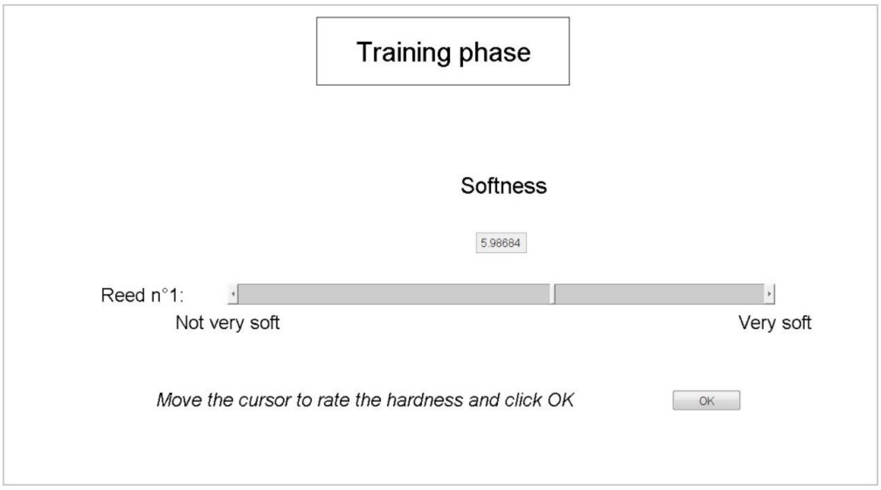
First, the training phase helped the subjects understand the meaning of the two descriptors Softness and Brightness. For that, we used “anchor stimuli”, which were located at the extremes of the scale under consideration. For Softness, the anchor stimuli were two reeds. One was considered to be as soft as the softest reed of the 20 in the product space, and one was considered to be as hard as the hardest reed. These extreme have been determined by several saxophonists who tested all the reeds and who agreed about which reeds were extreme. For Brightness, we used recorded sounds as anchors. To begin the training phase, the subjects tested anchor stimuli for each descriptor knowing which extreme was being presented. Then, participants tested the anchor stimuli again without knowing to which extreme it corresponded, and attempted to determine which one it was. Subjects with a success rate higher than 80% continued on, to the next step. This method is inspired from the training phase described in [4]. Finally subjects were asked to rate 3 quite different reeds on the interface they would use for the following phase in order to train them to use the interface.

After that training phase, the subjects began the real evaluation phase. During this phase, the subject used a Matlab interface to assess the reeds. The 20 reeds were proposed one at a time and the subject assessed them according to the descriptors Softness and Brightness on a continuous scale as shown in Figure 1. For global quality, a continuous scale was also proposed, but there were verbal anchors on the scale as shown in Figure 2. The reeds were presented to the subject in an order following a Williams Latin square. The presentation plan was perfectly balanced (see

Appendix A: Presentation plan of the reeds for the subjective tests). There were 2 repetitions and the subject could test again the anchor stimuli between the repetitions, if he needed to. The experimenter set the reeds on the mouthpiece for the players. We used two mouthpieces of the same model (a Vandoren V16 T7 Ebonite) in order to gain time (while the subject was assessing a reed, another one was set on the mouthpiece). To assess the reeds according to softness and brightness, the subjects were asked to play a few musical patterns as shown in Appendix B: Score of pattern presented for the subjective tests.

Finally, the subjects answered a short questionnaire asking them which mouthpiece, reed, saxophone and musical style they usually play and their past experience. The data of this questionnaire could eventually be useful to characterize some groups of subjects if they don't agree about some descriptors.

These subjective tests took place in a room at CIRMMT (Center for Interdisciplinary Research in Music Media and Technology) in McGill University, Montreal.



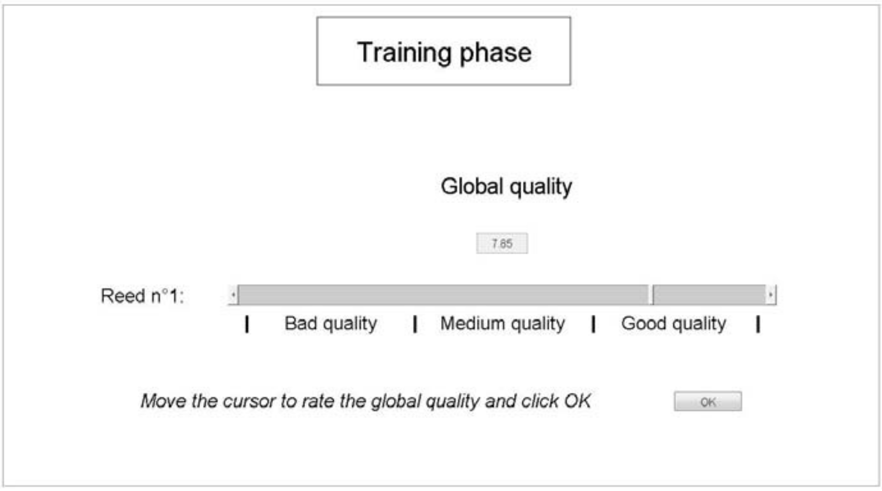
Training phase

Softness

Reed n°1: 5.98684

Move the cursor to rate the hardness and click OK

Figure 2: Continuous scale for the assessment of softness



Training phase

Global quality

Reed n°1: 7.85

Move the cursor to rate the global quality and click OK

Figure 3: Continuous scale for the assessment of the global quality

At the end of these tests, for each of the 10 subjects, we have 2 arrays of values (one per repetition). The arrays had 20 rows (one per reed) and 3 columns (one per descriptor) and contained the evaluation of the reed by the subject. In parallel with subjective test, we performed objective measurements on the reeds with other players.

2.2 Description of the in vivo measurements

2.2.1 Material

For the objective measurements, we chose to perform "in vivo" measurement. The principle is to perform measurements when the musician is playing the reed. The advantage is that we have a real playing situation, but this method has the disadvantage of introducing variability, particularly because of the way the musician play.

We chose to measure the acoustic pressure $\boldsymbol{p_a(t)}$ at the bell of the saxophone, the pressure in the musician's mouth $\boldsymbol{p_m(t)}$ and the pressure in the mouthpiece $\boldsymbol{p_{mp}(t)}$. The mouth pressure was measured using a differential pressure sensor Endevco 8507-C1 stuck in the mouthpiece. The pressure in the mouthpiece was measured by another Endevco 8507-C1 introduced into a hole drilled in the mouthpiece. And the acoustic pressure was measured by a B&K 4190-L-001 microphone placed in front of the saxophone bell. The sampling frequency used was 44100 Hz. A photo of the experimental device is presented in Figure 4 and an example of the measured signal is shown in Figure 5.

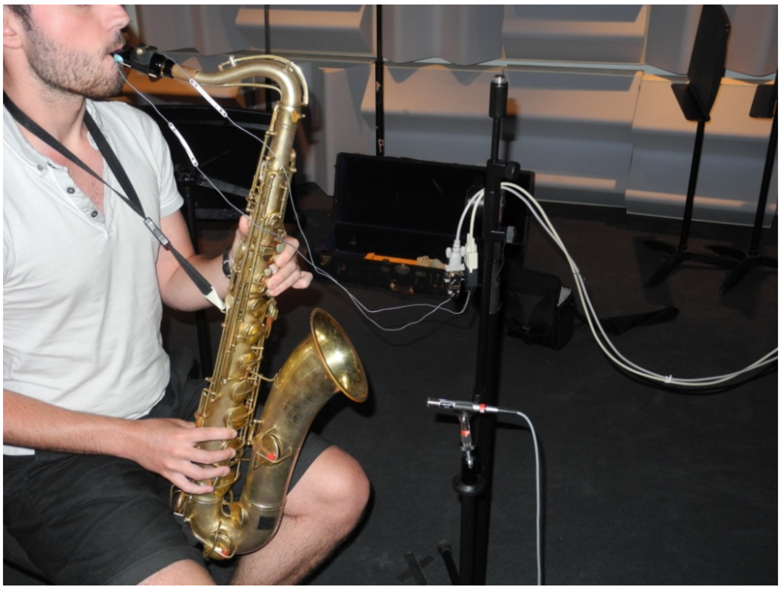


Figure 4: Photo of the experimental device.

Three saxophonists (PK, GS, BG) made the measurements for the 20 reeds with the same material as the subjective test concerning the mouthpiece and the reeds. The two players PK and GS performed two sessions of measurements two months apart and BG performed only one session. The pattern played by the saxophonists was an arpeggio of 7 notes (C4, G3, Eb3, C3, G2, Eb2, and C2)-concert key. But the playing of the seventh note (the lowest note: C2) was often of bad quality, so we chose to keep only the first six notes. This pattern was repeated 5 times for each reed and each saxophonist. The saxophone used by PK and GS was a Conn New Wonder and the saxophone used by BG was a Selmer Reference.

The measurement sessions for GS and PK took place in a lab at CIRMMT in Montreal, Canada. The first session was performed the 10th of May 2012 and the second one the 5th of July 2012. The measurement session of BG took place in the LAUM (Laboratoire d'Acoustique de l'Université du Maine) in LeMans, France the 11th of April 2012.

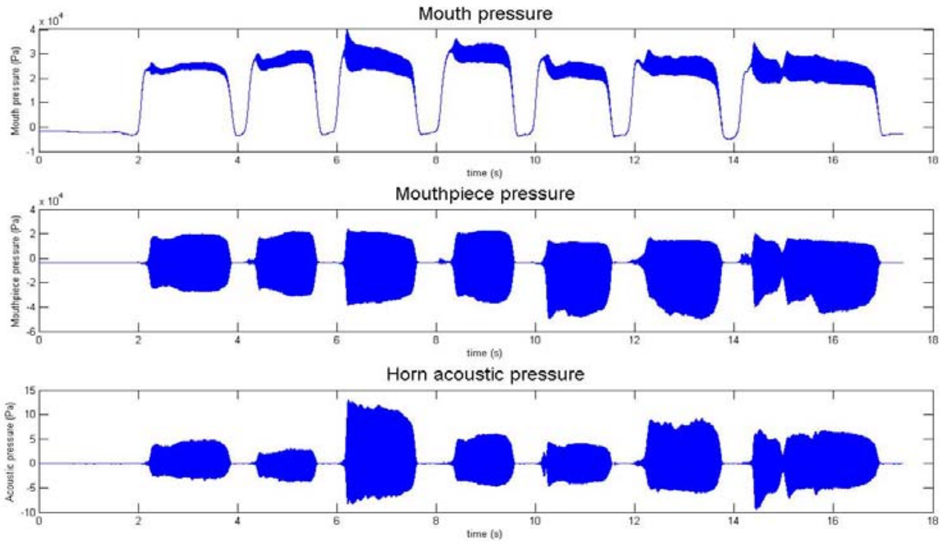


Figure 5: Example of "in vivo" measured signal.

2.2.2 Description of the "in vivo" descriptors

From this signal, we extracted several descriptors used in previous studies of B. Gazengel [3] [1]. Among them, several were acoustic descriptors computed from the harmonics of the spectral representation of the stationary part of the signal defined by calculating the energy of the acoustic signal $\boldsymbol{p_a(t)}$:

$$E(t) = \int_0^t {p_a}^2\,(\tau)d\tau$$

The stationary part of the signal is obtained for $E(t) \in [0{,}05; 0{,}95]E_{max}$ where E_{max} is the maximum energy obtained at the end of the note. To obtain our acoustic descriptor, we used 40 harmonics of the signal for each note. This number of 40 has been chosen to respect the Shannon condition. We have a sampling frequency $f_s = 44100\,Hz$ and the extreme frequency of the bandwidth we want to reach corresponds to the 40th harmonic of the higher note (C4: 523,25 Hz). This corresponds to a frequency of $40 * 523{,}5 = 20930\,Hz$, which is lower than $\frac{f_s}{2}$. The Shannon condition is thus respected.

Let us now present the acoustic descriptors we obtained, which were computed for each note and each reed. We consider that the notes are harmonics sounds, characterized by their spectrum in permanent regime, the frequency of the fundamental is f_1, the frequency of the harmonic k is f_k, and the amplitude of the harmonic k is A_k.

- The Spectral Centroid, which is a simple quantification of the distribution in the power spectrum, defined by :

$$SC = \frac{1}{f_1}\frac{\sum_{k=1}^{k=40} A_k f_k}{\sum_{k=1}^{k=40} A_k}$$

where f_k is the frequency and A_k is the amplitude of the kth harmonic

- The Odd Spectral Centroid (considering only the odd harmonics):

$$OSC = \frac{1}{f_1}\frac{\sum_{k=0}^{k=19} A_{2k+1} f_{2k+1}}{\sum_{k=0}^{k=19} A_{2k+1}}$$

- The Even Spectral Centroid (considering only the even harmonics):

$$ESC = \frac{1}{f_1}\frac{\sum_{k=1}^{k=20} A_{2k} f_{2k}}{\sum_{k=1}^{k=20} A_{2k}}$$

NB: All the spectral centroids are divided by the fundamental frequency in order to compare SCs of different notes.

- The ratio between odd and even harmonics:

$$OER = \sqrt{\frac{\sum_{k=0}^{k=19} {A_{2k+1}}^2}{\sum_{k=1}^{k=20} {A_{2k}}^2}}$$

- The amplitude of the harmonic pressure signal:

$$Lv = \sqrt{\frac{\sum_{k=1}^{k=40} {A_k}^2}{2}}$$

- The three Tristimulus components:

$$TR1 = \frac{{A_1}^2}{\sum_{k=1}^{k=40} {A_k}^2}$$

$$TR2 = \frac{{A_2}^2 + {A_3}^2 + {A_4}^2}{\sum_{k=1}^{k=40} {A_k}^2}$$

$$TR3 = \frac{\sum_{k=5}^{k=40} {A_k}^2}{\sum_{k=1}^{k=40} {A_k}^2}$$

- The “Tristimulus 4” corresponding to the ratio between the power of harmonics above 4000Hz and the total power of the harmonics.

$$TR4 = \frac{\sum_{k|f_{k>4000}} {A_k}^2}{\sum_{k=1}^{k=40} {A_k}^2}$$

After considering the stationary part of the signal, we took a descriptor from the transient part of the acoustic signal $p_a(t)$: the Attack Time. We first compute the envelope of the note by the convolution of a Hanning window and the absolute value of the signal. Then we define the Attack Time by:

$$AtT = t_e - t_b$$

Where t_b and t_e are defined as the times at which the envelope attains respectively 10% and 90% of its maximum value [5].

Afterwards, we used the signal of the pressure in the mouth to have two descriptors:

- The Mean Static Pressure is estimated as the mean of the pressure in the mouth during the stationary part of the signal:

$$StP = \frac{1}{t_{stat2} - t_{stat1}} \int_{t_{stat1}}^{t_{stat2}} p_m(\tau) d\tau$$

Where p_m is the pressure signal in the mouth and t_{stat1} and t_{stat2} are the beginning and the end, respectively, of the stationary part of the acoustic signal.

- The Pressure Threshold described in [2]. To compute it, we use a detection function D(t) on the acoustic signal $p_a(t)$ as follow:

$$D(t) = \sqrt{A(t)^2 + B(t)^2} \text{ with}$$

$$A(t) = \int_0^t p_a(\tau) \cos(2\pi f_1)\, d\tau$$

$$B(t) = \int_0^t p_a(\tau) \sin(2\pi f_1)\, d\tau$$

Then, with an empirical threshold, we deduce a time t_s of the beginning of the useful signal (Figure 6).

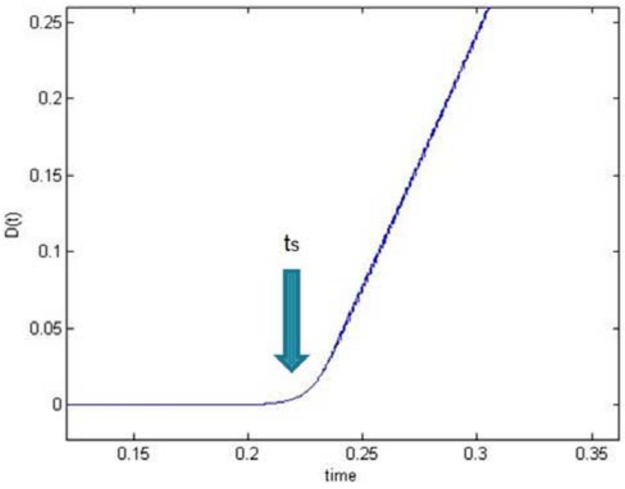


Figure 6: Detection function of the acoustic signal

The Pressure Threshold PTh is determined by finding the value of the mouth pressure at time t_s.

Finally we defined a last objective descriptor called efficiency, which is the ratio between the amplitude of the harmonic pressure signal and the mean static pressure:

$$Eff = \frac{As}{StP}$$

In conclusion, each reed is defined by 13 objective descriptors that are:

- The Spectral Centroid (SC)
- The Odd Spectral Centroid (OSC)
- The Even Spectral Centroid (ESC)
- The ratio between Odd and Even harmonics (OER)
- The amplitude of the harmonic signal (Lv)
- The 4 tristimuli (TR1, TR2, TR3, TR4)
- The Attack Time (AtT)
- The mean Static Pressure (StP)
- The Pressure Threshold (Pth)
- The efficiency (Eff)

All of these descriptors are computed for each note and each reed and each of the 5 repetitions of the pattern.

2.2.3 Processing of the recordings and calculation of the descriptors.

Once the recordings made, we have to “cut” them in order to isolate the repetitions and the notes. As a matter of fact, the recordings are made of the succession of 5 arpeggios of seven notes. The arpeggios were separated manually and for each repetition, the notes were separated by an automatic program. This program uses the energy $E(t)$ (see section 2.2.2) of the acoustic signal $p_a(t)$, to separate the notes by thresholding.

For all the acoustic descriptors, the amplitudes A_k are computed by taking the modulus value of the discrete Fourier transform of the acoustic signal $p_a(t)$. The program begins by isolate the stationary part of the signal still using the thresholding of the energy $E(t)$ of the signal. The thresholds used are the one described in the definition of the stationary part at the beginning of the section 2.2.2. After that, the program computes the discrete Fourier transform on this stationary part using the Matlab function: fft. The module of this transform is then computed. To access the value of the amplitude of the harmonics, we isolate an area of the spectrum where the harmonics theoretically should be and we take the maximum of this area and the corresponding frequency.

In conclusion, after calculate these harmonics we can compute all the acoustic descriptors.

3 Results of the subjective tests

To analyze the score given by the subjects during the subjective tests, we started by performing a consonance analysis on the data to study the consensus between the subjects.

3.1 Consonance analysis

The sensory panel consisted of J=10 assessors who judged I=20 products during K=2 sessions using M=3 attributes. So the assessment of product i by assessor j during session k according to descriptor m is denoted Y_{ijk}^{m}. Different tables of data can be formed. Let X_k^m denote the $(I$*$J)$ matrix describing the assessments made during session k on descriptor m by all the assessors.

The consonance analysis is based on PCA. The purpose of the consonance analysis is to estimate the agreement between the subjects in their evaluation of the reeds. A description of the method can be found in [6].

To study the agreement for each descriptor (independent of the sessions), the repetitions are merged vertically (repetitions are considered as different products). The PCA is made on the matrix $X^m(2\,I \times J)$:

$$X^m = \begin{bmatrix} X_1^m \\ X_2^m \end{bmatrix}$$

A perfectly consensual panel consists of assessors that use the descriptors and rate the reeds in the same way. So the more the panel is consensual, the more the arrows of the assessors go in the same direction. When the arrows go in totally different directions, we have a real disagreement. In this case, subjects should be divided into more consensual groups. The results (PCA of X^m) are given in Figure 7 for each descriptor. In this PCA, the variables are the subjects and the individuals are the reeds.

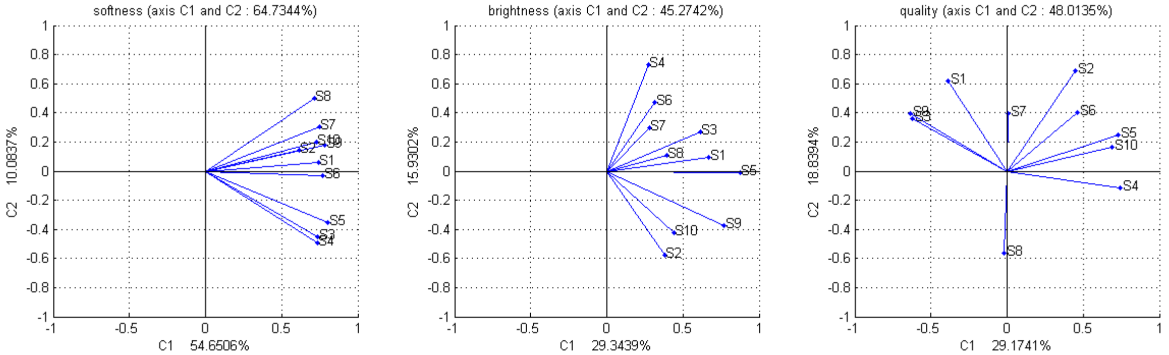


Figure 7: Consonance analysis for each descriptor: plot of 2 first factors of the PCA

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

$$C = \frac{\lambda_1}{\sum_{r=2}^{K} \lambda_r}$$

Where K is the component number in the PCA, 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	% Variance first PC
Softness	1,21	54,65%
Brightness	0,42	29,34%
Global quality	0,41	29,17%

Table 1: Results of consonance analysis for the panel subjects

In conclusion, 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 normal, given that this expresses preferences of the saxophonist, and that the tastes of the musician can be very diverse. Subjects S1, S3, S9 are rather opposite to the rest of the panel, subject S8 is independent of the general trend according to preference. We will probably have to analyze the quality separately from the two other descriptors and for several groups of subjects.

To confirm these conclusions we can also use other methods like the eggshell plot.

3.2 The Eggshell plot

This technique, described in [7], is interesting to visualize the differences between the subjects in ranking the objects. It is particularly useful to identify rankings that differ over all or just for part of the range of objects ranked.

The principle is to compute a consensus ranking, an underlying order (generally the first component of the PCA of the matrix of ranks), and to plot each assessor’s rank against the consensus. For each assessor, we compute the cumulative ranks according to the consensus order and after that, we subtract the cumulative ranks of a hypothetical assessor who would give the same note to all the reeds. The cumulative rank difference between the underlying order and the subject can then be plotted.

The consensus rank is given by a U-shape in the bottom of the graph. The disagreement between each assessor and the consensus is given by the “area” under the assessor’s polygon and the consensus U-Shape.

The eggshell plot for the 3 descriptors and the 20 reeds is given Figure 8. The results are of course in agreement with the consonance analysis. The consensus is stronger for softness.

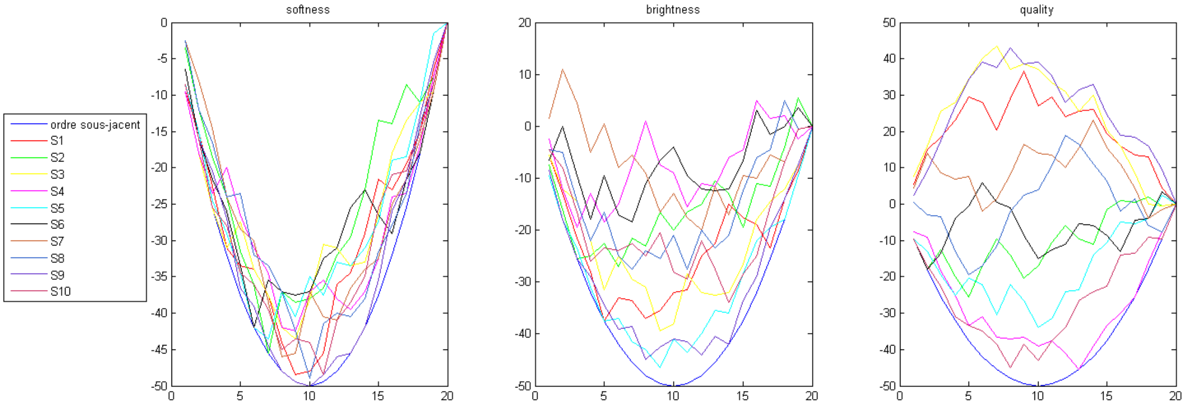


Figure 8: Eggshell plot for the three descriptors; x-axis the 20 reeds

3.3 Individual performances of the assessors

After studiying the quality of the consensus of the sensory panel, we can focus on the individual performances of the subjects to see if the results of some subject deserve to be discarded. We use in this section the principles of the GRAPES method [8], which provides a graphical representation of assessors' performances. We will focus on the different uses of the scale, the reliability of the subjects, their repeatability and their discrimination capacity.

3.3.1 Use of the scale

Two quantities can be computed to compare the use of scales by assessors. LOCATIONj is the average of the scores given by assessor j.

$$LOCATION_j = Y_{.j.}$$

N.B.: considering the evaluation Y_{ijk} (see section 3.1), the notation $Y_{.j.}$ means the mean of evaluations Y_{ijk} over the indices i (product) and k (session).

SPANj is the average standard deviation of a score given by assessor j within a session. It represents the average amplitude used by the assessor to discriminate the products.

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

The Figure 9 presents SPANj vs LOCATIONj for the different descriptors, and the different subjects S1 to S10.

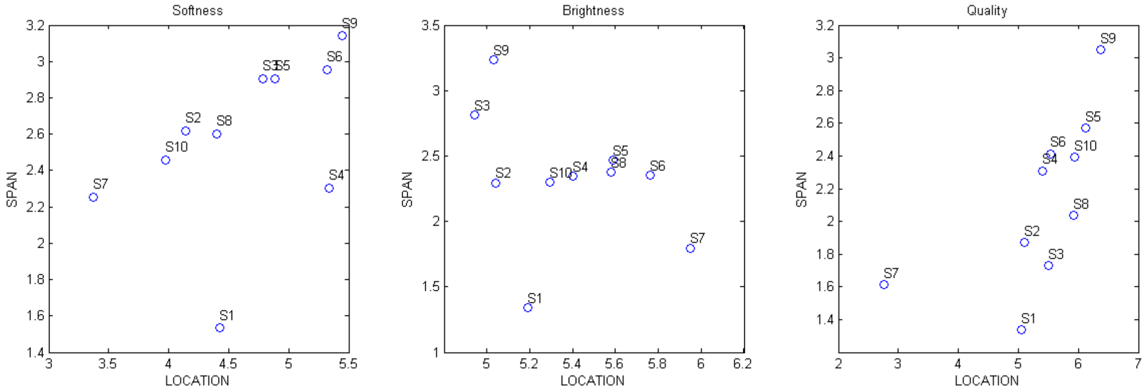


Figure 9: Plot of the SPANj vs LOCATIONj for each descriptor.

In conclusion, subject S1 uses a weak range for all the assessment (the SPAN is very weak) and subject S7 globally dislikes all the reeds, and assesses them as not soft (LOCATION is weak for this subject).

3.3.2 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, represents the measurement error of the subject, relative to the average amplitude used for the ratings. It is given by:

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

The DRIFT_MOOD is the between-sessions error relative to the average amplitude used for the ratings (i.e. expressed in SPAN units). It represents the deviation of the ratings of the subject across the sessions.

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

Figure 10 represents, for each descriptor, the performance of the subjects according to DRIFT_MOOD and UNRELIABILITY.

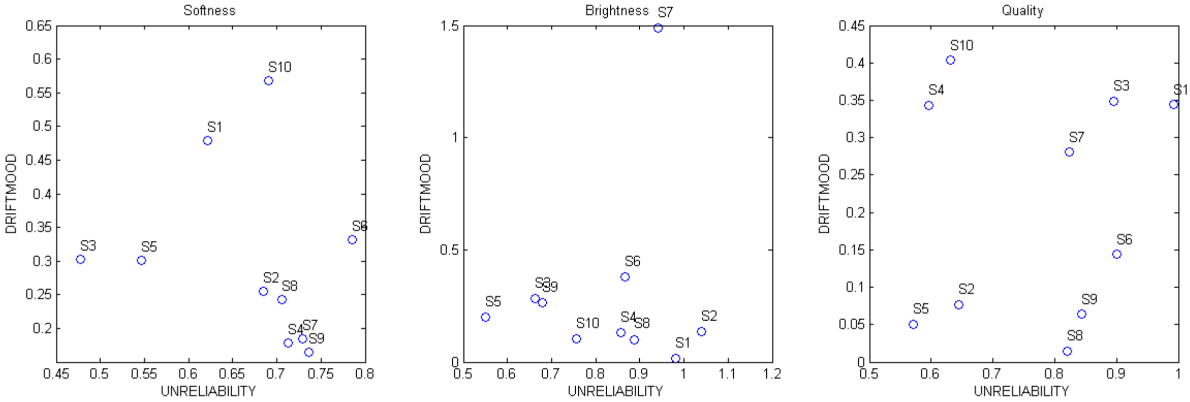


Figure 10: plot of the DRIFT_MOODj vs UNRELIABILITYj for each descriptor

In conclusion:

- For softness, S6 is the least reliable, 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, S5 is the most reliable. S7 deviates the most between the 2 sessions.
- For quality, S1 is the least reliable, 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.

3.4 Global performance of the panel

After the individual performances, we can focus on the global performance of the panel. We start by using the ANOVA model for the whole panel described before (equation (1)).

3.4.1 ANOVA

The principle of ANOVA is to model a dependent variable (the response) with independant variables (the factors)[1].

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 K, number of sessions

A model for the whole panel (equation (1)) can be created, taken into account the session effect, the product effect and also interaction the interaction between the session and the reeds:

$$\boldsymbol{Y_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + \epsilon_{ik}} \quad (1)$$

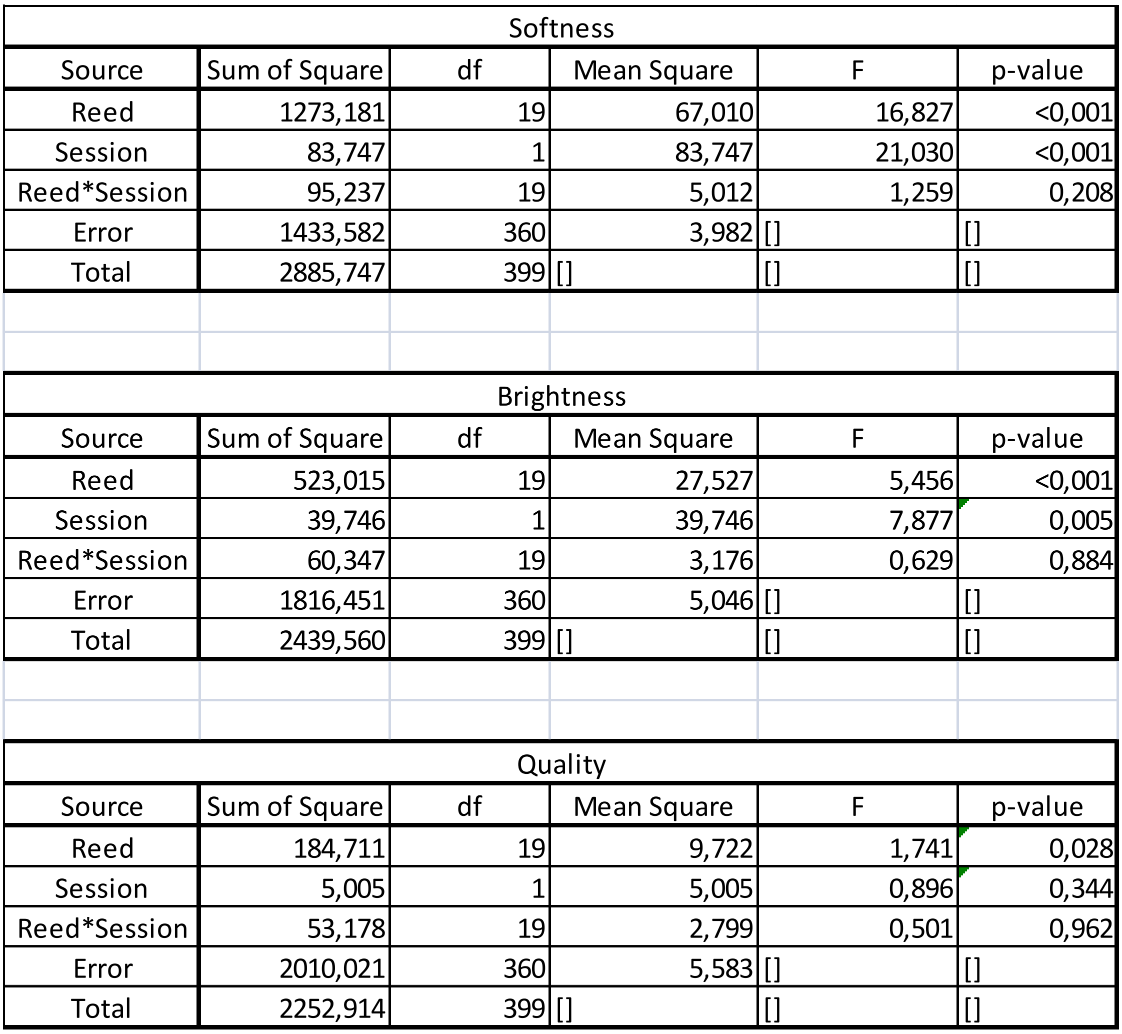
$\boldsymbol{\alpha_i}$: main effect of reed i
$\boldsymbol{\gamma_k}$: main effect of session k
$\boldsymbol{\alpha\gamma_{ik}}$: effect of the interaction session*assessor

Here we don't take the subject effect because we don't have enough values to estimate correctly the contribution of the subject effect, the reed effect, the session effect and the associated interactions in the same model, the degree of freedom of the residual would be to weak. So we consider only the reed effect, the session effect and the associated interaction that are the more important to us. As a matter of fact, the reed effect determines the discriminant power of the panel, and the interaction reed*session determines the repeatability of the panel. Consequently, the variable subject becomes a random variable in the model and gives us a bigger power of analysis. So we will have estimations that can be trusted.

A least square procedure is used to estimate the coefficients of the model. An ANOVA model is fitted for each descriptor.

[1] For all the factors of ANOVA (product, subjects), a fixed effect model is used. It signifies that the results cannot be generalized to the global population of subjects and products

The results of the ANOVA model for the whole panel (equation (1)) are given in Table 2.



Softness					
Source	Sum of Square	df	Mean Square	F	p-value
Reed	1273,181	19	67,010	16,827	<0,001
Session	83,747	1	83,747	21,030	<0,001
Reed*Session	95,237	19	5,012	1,259	0,208
Error	1433,582	360	3,982	[]	[]
Total	2885,747	399	[]	[]	[]

Brightness					
Source	Sum of Square	df	Mean Square	F	p-value
Reed	523,015	19	27,527	5,456	<0,001
Session	39,746	1	39,746	7,877	0,005
Reed*Session	60,347	19	3,176	0,629	0,884
Error	1816,451	360	5,046	[]	[]
Total	2439,560	399	[]	[]	[]

Quality					
Source	Sum of Square	df	Mean Square	F	p-value
Reed	184,711	19	9,722	1,741	0,028
Session	5,005	1	5,005	0,896	0,344
Reed*Session	53,178	19	2,799	0,501	0,962
Error	2010,021	360	5,583	[]	[]
Total	2252,914	399	[]	[]	[]

Table 2: Table of ANOVA for the three descriptors softness, brightness and quality

We see here that the reed effect is significant for all the descriptor which is good. It means that the whole panel well discriminated the reeds. We also see that the interaction reed*session is not significant for all the descriptor. It means that the assessments of the panel are in agreement form a session to another which is a good thing too.

So we can use now the assessments made by the panel to consensual values of the descriptors for the reeds by performing a multivariate analysis of the assessments.

3.5 Multivariate analysis of the assessments

The agreement among the assessors of the panel is an important problem in sensory analysis. The direct mean value of the assessments of all the subjects may lead to a poor description of the differences among products if the subjects are not in agreement and have not a normal distribution (the mean value would not be representative in such a case). The sensory analyst is confronted with the dilemma of discarding dissonant assessors, and losing in this case information, or leaving the data as such, and getting a noisy assessment.

Several methods are proposed to transform the individual evaluations in an average multivariate description of the products.

N.B: the descriptor "Quality" is excluded from this analysis because it is too subjective by nature and no agreement is required with preference in sensory analysis.

The first method is to compute the average values of the reeds according to the 2 descriptors Softness and Brightness, for the 10 subjects, denoted $Y_{i..}$. The representation is given Figure 11.

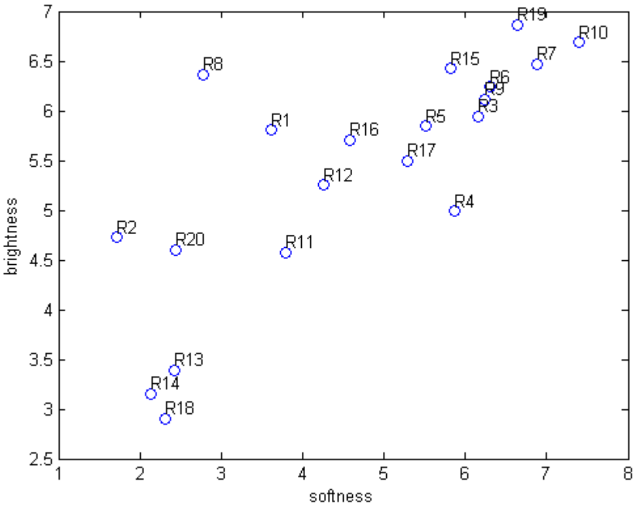


Figure 11: Position of the reeds according to the softness and the brightness (average configuration)

- R10 , R7, R19 are the most soft and bright reeds
- R14, R18, R13 are the least soft and bright reeds

We also can point out a correlation between the two descriptors Brightness and Softness: a bright reed is also generally soft.

To get a group average configuration, we used several other methods of multivariate analysis like the GAMMA method (see Appendix D: The GAMMA method) or the Generalized Procrustes Analysis (see Appendix E: Generalized Procrustes Analysis) to obtain mean values. These methods did not provide more information and gave finally similar results than the simple average configuration. So we decided to use the scores of the average configuration to characterize the reeds for the rest of this report.

3.6 Analysis of global quality

For the attribute "Quality", consider the matrix X_{123}^{m} of dimension (2I×J), which considers repetitions as additional variables (variable = subject*session).

$$X_{123}^{m} = (\bar{X})$$

A cluster analysis with Hierarchical Ascendant Classification can be made on the matrix X_{123}^{m}. We performed the cluster analysis on the row data because of the anchored scale (Figure 3). This verbal anchoring of the scale gives a meaning to the scores and the mean. So if we centered and reduced the data, we would lose this meaning. The distance used for the HAC is the Euclidian distance, the linkage rule is the Ward criterion (variance criterion). The dendrogram of the classification is presented in Figure 12.

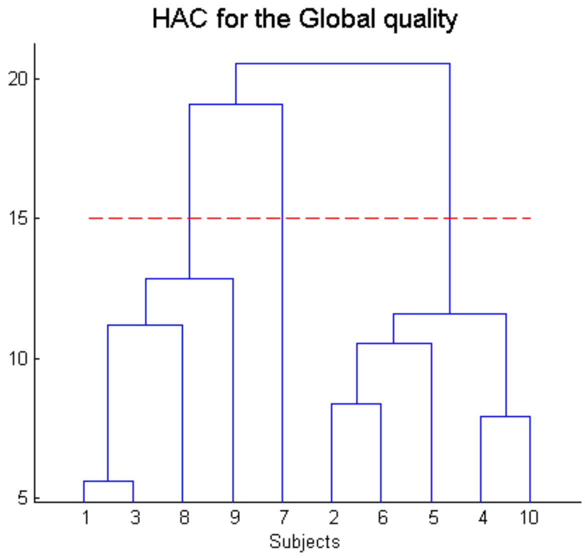


Figure 12: Dendogram of the HAC according to the global quality ratings for the mean of the 2 sessions

3 clusters can be formed:

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

The scores of the reeds for the two groups 1 and 2 are given in Figure 13.

Group 1 and 2 have mainly conflicting opinions on reeds R13 and R18.

The typical features of the classes are:

- Group 1 (typical subject S3) appreciates R13 and rejects R18.
- Group 2 (typical subject S10) appreciates R18 and rejects R13.

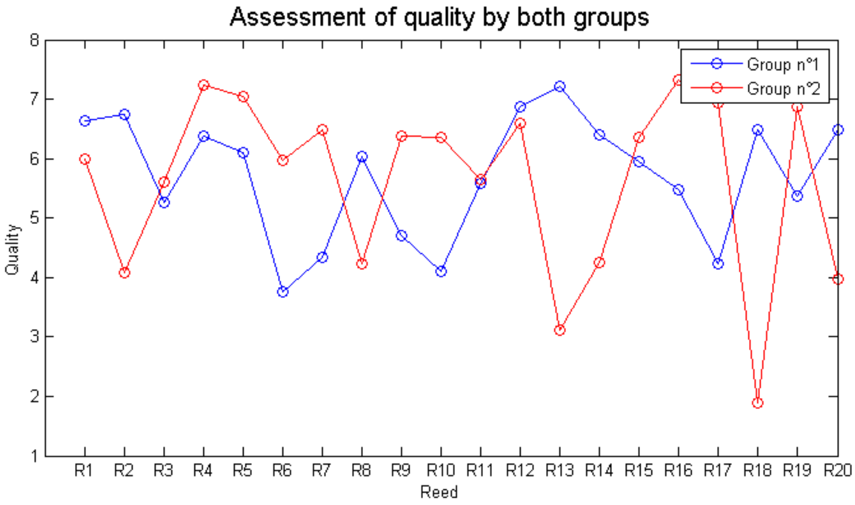


Figure 13: preference scores for the 2 different groups

We tried to characterize both groups with information concerning the subjects obtained from the questionnaires, but no feature of the musicians seems to characterize the groups. And the small number of subjects doesn't help. 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 in Figure 13 between the two groups are on the softest reeds or on the hardest reeds. For example we can see big differences for the reeds 2, 13 and 18 which are perceived as the hardest reeds, and we also see big differences for the reeds 10 and 17 which are perceived as soft reeds.

Despite this lack of characterization, we will use this data of the global quality to try to produce a predictive model for each group of subject.

4 In vivo measurements of the reeds

4.1 Descriptive statistics

After analyzing the results of the subjective tests, we need to analyze the results of the objective measurements. It would be too fastidious to represent all the values because of their high number. As a matter of fact, there are 20 reeds*5 repetitions*6 notes*13 variables*3 musicians and *2 sessions for PK and GS. To visualize the variance of the data and the differences among musicians, we chose to represent only a few variables for one note, the two sessions for GS and PK. We will do as if there are 5 musicians for an ease of representation of the different sessions (PK1, PK2, GS1, GS2, BG). In the Figure 14 is the representation of the Spectral Centroïd of the 20 reeds, for each musician, the note G3. First, we want to see if taking the mean value of the descriptors over the 5 repetitions makes sense. To give an example, in Figure 14 are presented the boxplots of the Spectral Centroïd and the Pressure Threshold for the note G3 and the player PK1. These boxplots take the values of the 5 repetition for each descriptor. All the boxplots for all the descriptors, for all the players and for the note G3 are presented in the annex.

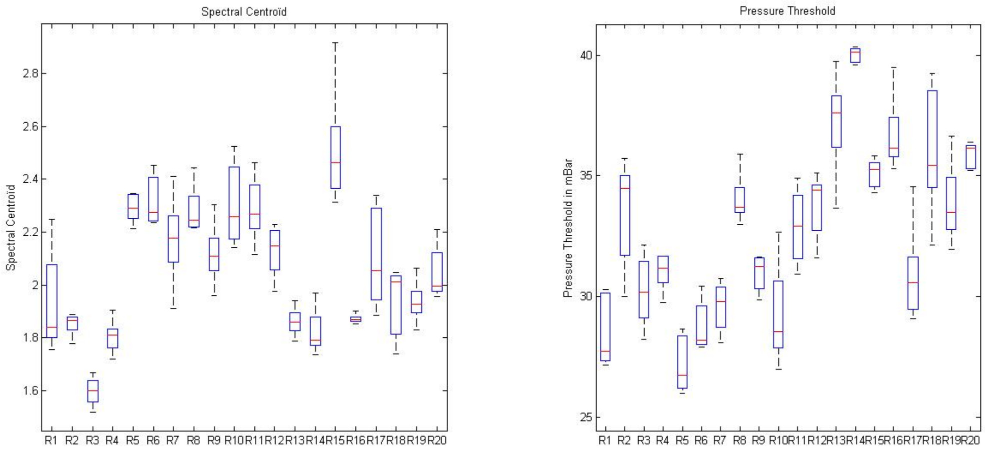


Figure 14: Boxplots of the 5 repetitions for the 20 reeds and for the descriptors SC and PTh

First we can see that the variance is irregular depending on the reed. Some reeds have a high variance like the reed 17 for the spectral Centroïd or the reed 18 for the Pressure threshold. But even though the repetition error can be high for some reeds, we see that this repetition error is low for most of the reeds and that the reeds are well discriminated by these two descriptors. So taking the mean value of the descriptors over the 5 repetitions makes sense.

After taking the mean value over the 5 repetitions, we can study the influence of the note on the descriptors. In the Figure 15 are presented the plots of the mean value of the descriptors SC and PTh over the 5 repetition, for the player PK1. There is one curve by note.

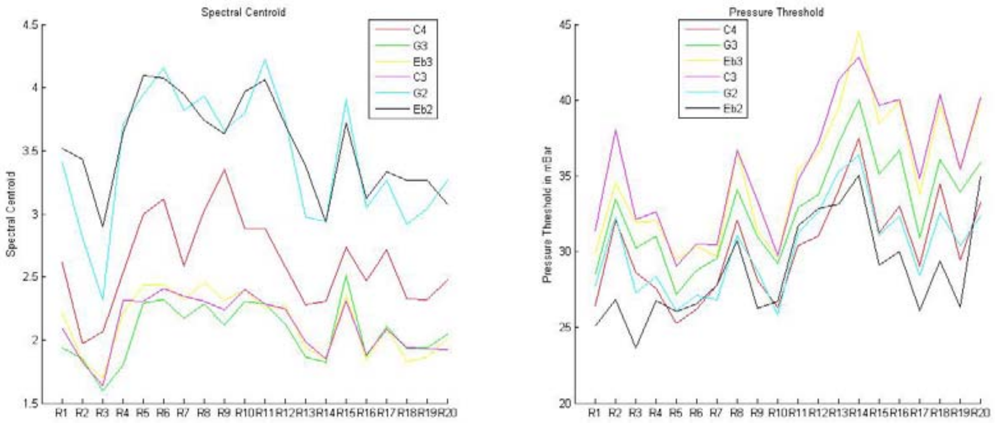


Figure 15: Plots for each note of the mean value of the descriptor SC and PTh over the 5 repetition for the player PK

We see in Figure 15 that the note has an influence on the values of the descriptors. But, we also see that the curves have the same evolution. So the average curve over the 6 notes will have the same evolution too. So taking the mean value over the 6 notes makes sense.

After studying the influence of the note, we can look at the variance of the values of these two descriptors for all the musicians. In Figure 16 is the representation of the boxplot of the 5 musiscians for the descriptors SC and PTh, the note G3.

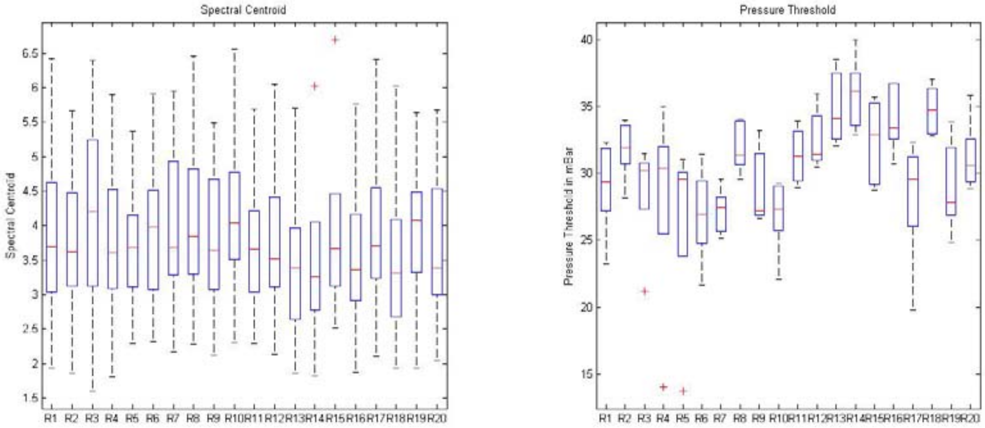


Figure 16: Boxplots of the 5 musicians for the 20 reeds and for the descriptors SC and PTh

Here we can see that for the Spectral Centroid, the variance is high for all the reeds. The reeds can't be distinguished according to this descriptor. So we can say there is a problem given the fact that the reeds can be distinguished when we took the musicians separately. This is probably due to the lack of agreement between the musicians according to the Spectral Centroid. The things are different for the Pressure Threshold. We see that the intra-reed variance is lower than for the SC and that the reeds can be discriminated. For this descriptor, the musicians seem to have a higher agreement. To precise this problem of agreement, we performed a PCA on the same data to analyze the consensus between them. In Figure 17 are presented the PCAs made on the matrix (musicians*reeds) for the two descriptors SC and PTh. The musicians are here the variables and the reeds are the individuals.

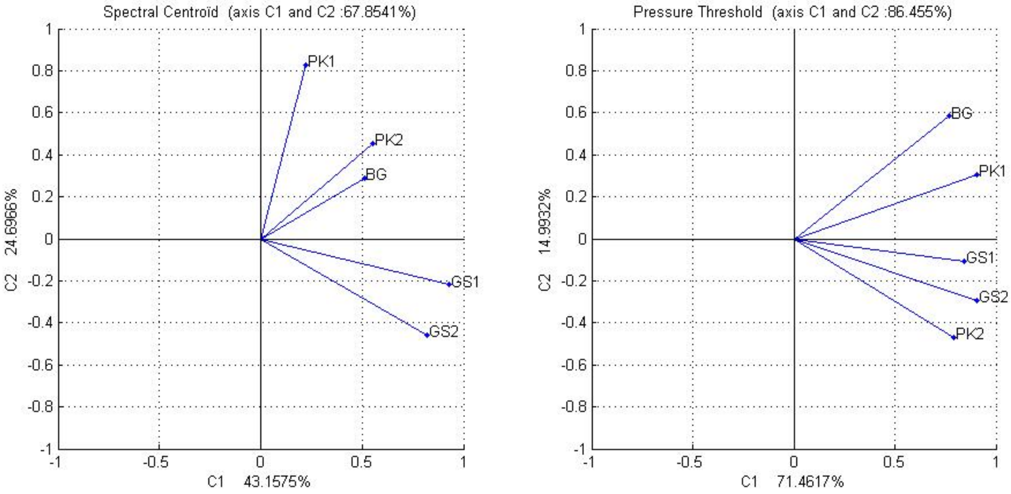


Figure 17: PCA of the values of the descriptors SC and PTh for the 5 musicians

We can see here what we supposed before. For the Pressure Threshold, the agreement between the musicians is good and the first principal component gathers 71% of the total variance. For the Spectral Centroid, we see that consensus is much lower than for the Pressure Threshold. But we can see that the agreement is better between the two sessions of the same musician than between two different musicians. For example, we see that the arrow of GS1 is closer of the arrow of GS2 than the other musicians. For some descriptors like the Spectral Centroid, we see that there is variability between the musicians. This variability among the measurements may be due to the way of playing of the players etc…

In conclusion, if we want to use all of these data as objective measurements, we need to deal with variability and keep only the information concerning the reed by fitting statistic models or by extracting consensus.

4.2 Individual results of the musicians

Let's begin by studying the individual results of the musicians.

4.2.1 One-way ANOVA model

To study the individual performances of the musicians for the objectives measurements, we choose to fit an ANOVA model to the data for each subject and variable labeled Z in equation (2). This model takes into account the reed effect. We don't take into account the note effect for the same reason as in the section 3.4, we want to have a better power of analysis. We use one model for one session (equation (2)). As the musicians GS and PK had participated two sessions of measurements, we will proceed as if there were 5 saxophonists: GS1, PK1, GS2, PK2, and BG. For all the considered models in the following, Z_{ijk} is a generic notation that represents a value of an objective descriptor for the ith reed, the jth musician and the kth session.

$$(2) \quad \boldsymbol{Z_{ijk} = \mu + \alpha_i + \epsilon_i}$$

α_i : main effect of reed i

For each musician and each descriptor, Table 3 presents the p-values of the test of the reed effect (coefficients α_i). In green are the significant effects.

P-values of the reed effect for each descriptor and player						
		Players				
		GS1	PK1	BG	GS2	PK2
Descriptors	AtT	<0,001	<0,001	0,111	<0,001	<0,001
	SC	<0,001	<0,001	0,892	<0,001	0,620
	OSC	<0,001	<0,001	0,818	<0,001	0,772
	ESC	<0,001	<0,001	0,966	<0,001	0,257
	OER	0,979	0,233	0,995	0,839	0,023
	Lv	0,014	<0,001	0,147	0,835	0,052
	TR1	0,984	<0,001	0,557	0,503	<0,001
	TR2	0,976	0,003	0,754	0,985	0,102
	TR3	<0,001	<0,001	0,454	0,056	0,972
	TR4	<0,001	<0,001	<0,001	<0,001	0,001
	PTh	<0,001	<0,001	<0,001	<0,001	<0,001
	StP	<0,001	<0,001	<0,001	<0,001	<0,001
	Eff	<0,001	<0,001	<0,001	<0,001	<0,001

Table 3: p-value of the reed effect for ANOVA model for the objective measurements.

First we see that for PK1 and GS1, there is a significant product effect for most o the descriptors. So in the first session, for PK and GS, the reeds were discriminated for most of the descriptors. In the second session we see we have less significant effects. But there is still a majority of the descriptors that discriminated the reeds. On the other side, we see that BG has few descriptors that have a significant effect. The measurements of BG are globally not discriminant. This, added to the fact that the measurements of BG were performed in a different place (different country), with a different saxophone and with a different material, we prefer to discard the measurements of BG for the following (there is a lot of parameters whose we can't know the influence on the measurements).

4.2.2 Two-way ANOVA model

For the musicians PK and GS, we study an ANOVA model that took into account the effect of session (equation (3)). Again, we examine the product effect to verify their discriminant power. The results are presented in the Table 5.

$$(3)\quad \boldsymbol{Z_{ijk} = \mu + \alpha_i + \gamma_k + \epsilon_{ij}}$$

α_i : main effect of reed i

γ_k: main effect of session k

P-values of the reed effect for each descriptor and player			
		Players	
		GS	PK
Descriptors	AtT	<0,001	<0,001
	SC	<0,001	<0,001
	OSC	<0,001	0,002
	ESC	<0,001	<0,001
	OER	0,639	0,243
	Lv	0,130	<0,001
	TR1	0,369	<0,001
	TR2	0,852	0,008
	TR3	<0,001	0,729
	TR4	<0,001	0,001
	PTh	<0,001	<0,001
	StP	<0,001	<0,001
	Eff	<0,001	<0,001

Table 4: p-value of the reed effect for the three-way ANOVA model for the objective measurements.

The conclusion is the same as previous: the two subjects have significant effects for most of the objective descriptors. The two subjects are globally discriminant: they see differences between the reeds. The only variable we can doubt is the ratio between the odd and even harmonics because he has no significant effect for any of the musician. However, the variable had in the previous part a significant effect for PK2, so we choose to keep this descriptor in the following.

We can now study, the correlation between the red effect coefficients α_i of GS1, PK1, GS2, PK2 for all the descriptors to see if their variability over the reeds is the same.

4.3 Inter-musicians and inter-sessions correlation

To see the inter-musician and inter-session correlation, we compute the Spearman correlation coefficient between the reed effect coefficients α_i for each player extracted from the previous ANOVA model (equation (4)). Table 5 presents the results for each pair of players for session 1 and 2 and the correlation between session 1 and 2 for the two players in order to assess the inter-session agreement. In green are the values of coefficients who are higher than 0.6.

Spearman coefficient for each pair of players					
		Players			
		PK1-GS1	PK2-GS2	PK1-PK2	GS1-GS2
Descriptors	AtT	0,59	-0,14	0,53	-0,09
	SC	0,25	0,25	0,20	0,90
	OSC	0,19	0,13	0,25	0,89
	ESC	0,20	0,54	0,37	0,84
	OER	-0,05	0,63	-0,34	0,37
	Lv	0,40	0,22	0,23	0,12
	TR1	0,47	0,01	-0,12	0,69
	TR2	0,19	0,30	-0,19	0,48
	TR3	0,40	0,17	0,30	0,84
	TR4	0,13	0,20	0,10	0,76
	PTh	0,82	0,83	0,75	0,84
	StP	0,78	0,53	0,09	0,59
	Eff	0,65	0,31	-0,15	0,41

Table 5: Inter-musician and inter-session correlation.

We see in Table 5 that the correlation between players (PK1-GS1, PK2-GS2) is very low except for Pressure Threshold (PTH) that has a strong correlation in all cases. This descriptor is computed from the mouth signal. So we see that the players generally have low correlations on the descriptors relative to the acoustic signal, their variability over the reeds is not the same. We also see that PK seems to be not repeatable because the correlation between both sessions is low (weak general correlation PK1-PK2). But GS (GS1-GS2) gets strong correlations for several descriptors. So GS seems to be the most consistent and the most repeatable musician.

As for the subjective data, we are facing the dilemma to discard musicians who are not enough reliable (and to lose perhaps a useful information), or to keep all the data (and to get possibly inconsistent data). In conclusion, we see that there is a lot of variability in the objective measurement. If we want to keep all the data, we have to find a way to extract from these data, consensual values that come from the contribution of all the players.

4.4 Definition of consensual measurements

4.4.1 Fitting of an ANOVA model

We saw that there is variability in the objective measurements. This variability can possibly be due to the added contributions of the player, the reed, the note and the session. In our case, we want to keep only the contribution of the reed. So we have to find a way to separate these contributions and have consensual measurements about its contribution. To obtain consensual measurements using the data of PK and GS, we used an ANOVA model to keep the essential common information about the reeds and remove useless variability caused by the change of player, the different sessions, the interaction between the musician and the reed, etc… The reed effects obtained will be the final objective measurements. The objective of the model is to filter the data and remove the possible unnecessary error (grasp the useful part of the information in the ANOVA coefficient). We have to keep a significant reed effect for all the descriptors to have values which make sense.

We build an ANOVA model taking account of the reed effect, the session effect and the musician effect. To take account of the effect of the interaction between the musician and the reed and the interaction between the musician and the session, we use a model with interaction as following (equation (4)). We consider the interaction between the musician and the reed because obviously, the musicians don't agree between them and this interaction is supposed to model this disagreement. We also considered the interaction between the musician and the session because for the same musician, some evaluations seem to change from a session to another and as previous, this interaction is supposed to model this change. We didn't take account of the interactions reed*session and reed*session*musician because after studying a complete model, it appeared they were not significant for most of the descriptors.

$$\boldsymbol{Z_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + \epsilon_{ijk}} \quad (4)$$

α_i : main effect of reed i
β_j: main effect of musician j
γ_j: main effect of session k
$\alpha\beta_{ij}$: interaction effect reed/musician
$\beta\gamma_{jk}$: interaction effect session/musician

In Table 6 are presented the p-values of the product effect for all the objective descriptors.

		p-value of reed effect
Descriptors	AtT	<0,001
	SC	<0,001
	OSC	<0,001
	ESC	<0,001
	OER	0,14
	Lv	<0,001
	TR1	<0,001
	TR2	0,01
	TR3	<0,001
	TR4	<0,001
	PTh	<0,001
	StP	<0,001
	Eff	<0,001

Table 6: p-value of the reed effect for the global ANOVA model

We see that the reed effect is significant for all the descriptors except for the OER. The reeds are discriminated by the global model. So using the values α_i of the reed effect makes sense for the correlation part. But fitting a statistic ANOVA model on the data to remove the variability is not the only option. We can also extract the consensus between the objective measurements by using the GAMMA method described in the Appendix D: The GAMMA method.

4.4.2 GAMMA method to define the consensual configuration

Method

Instead of fitting a statistic model like the ANOVA, we can try to extract the consensus between the measurements using the GAMMA method. This method makes the average value of the assessments of objective descriptors, taking into account weights for each musician and session, according to their agreement with the whole group[9]. It is based on a fixed vocabulary, common for all the musician. The interpretation of the common configuration is in this way easy, because it corresponds directly to the descriptors used by the subjects. It consists in computing weights for each subject, which are representative of the degree of agreement of the subject with the rest of the panel.

Let X_i denote the (I×M) matrix, describing the assessments made by the musician i on the I reeds according to M objective descriptors. The values of X_i are the average of the descriptors over the 6 notes and the 5 repetitions considered in the “in vivo” measurements. Matrix X_i is called a configuration. The first step of the method is, for each configuration, to center (removing the effect of the judge on the scoring) and rescale (standardize the data to the same total variance) the data.

- center the data : substract, for each data, the average of the column of the X_i matrix. This will give matrix Xc_i.
- standardization: multiply all the data by the factor $\theta_i = \frac{1}{\sqrt{trace(Xc_i^t.Xc_i)}}$. This will give matrix Y_i.

To assess the similarity between two subjects i and j, the method computes the following quantity that is directly the correlation coefficient between Y_i and Y_j, considered as vectors by rearranging the data in a single column.

$$t_{ij} = \frac{trace(Y_i^t.Y_j)}{\sqrt{trace(Y_i^t.Y_i)}.\sqrt{trace(Y_j^t.Y_j)}}$$

The similarity matrix S_{ij} of size (J×J) between all the assessors if given by:

$$S_{ij} = \frac{1+t_{ij}}{2}$$

The eigenvector $\beta = [\beta_1, \beta_2, \dots, \beta_J]^t$ associated with the largest value, represents the degree of agreement of the subjects with the rest of the panel. A weighted average configurations is finally computed to characterize the products:

$$C = \sum_{j=1}^{J} \beta_j.Y_j$$

C is next subjected to a PCA to characterize the differences between products.

To characterize the performance of the panel, two indices are computed,

$$\alpha_i = \frac{trace(Y_i^t.C)}{\sqrt{trace(C^t.C)}} \text{ ,the correlation coefficient between Yi and C}$$

$$\gamma = \frac{1}{J}\sum_{j=1}^{J}\alpha_j$$: the average correlation coefficient for the panel (performance index)

Results

With the whole group of 4 musicians under consideration, the values of the alpha coefficient for the average value across the sessions are given in Figure 18.

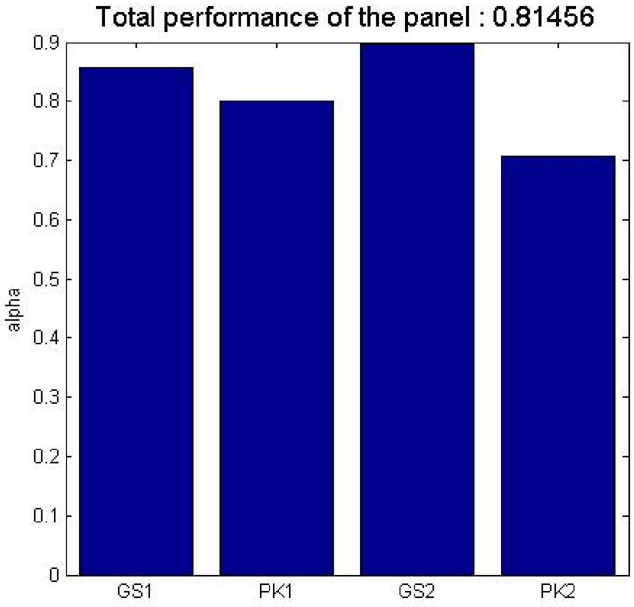


Figure 18: plot of the alpha coefficient for the whole group of 5 musicians

The performance index of the panel is 0.79, which is high. Finally the consensus between the players seems to be high despite the variability we observed previously. Now, we will try to build predictive models with the data extracted from the ANOVA models and from the gamma method and to compare them.

5 Correlation between subjective and in vivo measurements

5.1 One-to-one correlation

5.1.1 With ANOVA coefficients

For the correlation between subjective and in vivo measurements, we started by using the tool the most used in the literature: the simple one-to-one correlation using the Pearson coefficient of correlation. In Table 7 are presented the correlation coefficients between the coefficients of reed effect α_i of the global model with interaction (equation (4)) for the 13 objective variables on the one hand, and the average values of the reeds according to Softness, Brightness and Quality ratings of the two groups of Quality for the subjective descriptors, on the other hand. In green are the values greater than 0.8.

		Softness	Brightness	Quality grp 1	Quality grp 2
Descriptors	AtT	-0,57	-0,33	0,57	-0,56
	SC	0,64	0,84	-0,55	0,46
	OSC	0,62	0,83	-0,53	0,44
	ESC	0,70	0,84	-0,57	0,50
	OER	-0,19	0,04	0,02	-0,13
	Lv	-0,38	-0,39	0,33	-0,22
	TR1	-0,24	-0,42	0,36	-0,31
	TR2	-0,01	0,10	-0,13	0,12
	TR3	0,59	0,80	-0,60	0,49
	TR4	0,37	0,45	-0,34	0,14
	PTh	-0,82	-0,81	0,75	-0,59
	StP	-0,67	-0,80	0,65	-0,58
	Eff	0,23	0,33	-0,33	0,27

Table 7: Correlation coefficients between subjective and objective descriptors with the values of anova coefficients.

First, we can see that for the global quality, the correlation coefficients for both groups of subjects are not so high. Despite this, we can point out that all descriptors for group 1 are inversely correlated with the values of group 2, which suggests that the two groups have different evaluations of the reed quality. For example, for the Pressure Threshold, group one have a positive correlation (0.75), and group 2 have a negative one (0.59). As seen in the section 3.6, group 1 seemed to prefer hard reeds and group 2 seemed to prefer soft reed. So from a physical point of view, these correlations make sense because a “soft” reed necessitates a low pressure and a “hard” reed a high pressure.

Then, for Softness, only the Pressure Threshold has a strong correlation (-0.82). This negative correlation makes sense from a physical point of view: a “soft” reed necessitates a low pressure and a “hard” reed a high pressure.

Brightness has a strong correlation (-0.81) with the Pressure Threshold, the mean Static Pressure (-0.80), the Tristimulus 3 (0.80), the Odd Spectral Centroïd (0.83), the Even Spectral Centroid (0.84) and finally with the Spectral Centroid (0.84) which is in agreement with the literature. These correlations make sense too from a physical point of view: a “bright” reed will produce a sound with a high Spectral Centroid and a “dark” reed will produce a sound with a low Spectral Centroïd.

But the correlation coefficients are not high enough to build an accurate predictive model from a regression based on only one objective variable. So we chose to explain the subjective descriptor by more than one variable.

5.1.2 With the values from GAMMA method

Here we performed the same thing as previous but by using the values from the GAMMA method. In Table 8 are presented the correlation coefficients between values from the GAMMA method computed in the section 4.4.2 for the 13 objective variables on the one hand, and the average values of the reeds according to Softness, Brightness and Quality ratings of the two groups of Quality for the subjective descriptors, on the other hand. In green are the values greater than 0.8.

		Softness	Brightness	Quality grp 1	Quality grp 2
Descriptors	AtT	-0,78	-0,55	0,78	-0,62
	SC	0,56	0,76	-0,40	0,47
	OSC	0,52	0,73	-0,37	0,46
	ESC	0,68	0,81	-0,48	0,52
	OER	-0,36	-0,17	0,20	-0,18
	Lv	-0,28	-0,28	0,24	-0,11
	TR1	-0,07	-0,23	0,21	-0,25
	TR2	-0,13	0,04	-0,09	0,08
	TR3	0,36	0,43	-0,30	0,39
	TR4	0,41	0,46	-0,29	0,20
	PTh	-0,83	-0,82	0,78	-0,59
	StP	-0,68	-0,80	0,70	-0,56
	Eff	0,08	0,17	-0,19	0,21

Table 8: Correlation coefficients between subjective and objective descriptors with the values of the GAMMA method.

For Global Quality, conclusions are the same as before. The correlations are not strong but we still observe the opposition between the two groups regarding the sign of the correlation.

For Softness, it also the same as before, the Pressure Threshold has a strong correlation (-0.83).

For Brightness, there are less objective variables that have strong correlation with Brightness. But the three objective variables that have a strong correlation with Brightness had strong one previously. These variables are the Pressure Threshold (-0.82), the mean Static Pressure (-0.80) and the Even Spectral Centroid (0.81). The Spectral Centroid and the Odd Spectral Centroid have strong correlation too but lower than before (respectively 0.76 and 0.73).

The global conclusion is the same as before: the correlation coefficients are not high enough to build an accurate predictive model from a regression based on only one objective variable. So we chose to explain the subjective descriptor by more than one variable.

5.2 Regression model

The first thing that comes to mind is to consider several explanatory variables to build a predictive model with multiple linear regression. But the linear model suffers from issues regarding the multicollinearity of variables, which is here the case. All our objective descriptors are more or less correlated. Table 9 presents the correlation matrix of the objective descriptors (coefficients $\alpha_i$of the global model with interaction), in green are the values higher than 0.8. We can see that many descriptors are highly correlated. So if we use the multiple linear regression, the variance of the estimates of the coefficients may be very high because of multicollinearity issues as shown in [10].

	AtT	SC	OSC	ESC	OER	Lv	TR1	TR2	TR3	TR4	PTh	StP	Eff
AtT	1,00	0,03	0,04	-0,03	0,13	-0,10	0,07	-0,06	-0,05	0,08	0,18	0,12	-0,30
SC	0,03	1,00	1,00	0,98	0,14	-0,53	-0,41	0,05	0,88	0,71	-0,90	-0,84	0,18
OSC	0,04	1,00	1,00	0,96	0,16	-0,52	-0,43	0,07	0,90	0,70	-0,89	-0,85	0,21
ESC	-0,03	0,98	0,96	1,00	-0,01	-0,64	-0,25	-0,11	0,80	0,77	-0,90	-0,83	0,03
OER	0,13	0,14	0,16	-0,01	1,00	0,32	-0,63	0,65	0,26	-0,06	0,00	-0,12	0,54
Lv	-0,10	-0,53	-0,52	-0,64	0,32	1,00	-0,37	0,66	-0,38	-0,70	0,58	0,62	0,58
TR1	0,07	-0,41	-0,43	-0,25	-0,63	-0,37	1,00	-0,91	-0,63	0,20	0,42	0,42	-0,89
TR2	-0,06	0,05	0,07	-0,11	0,65	0,66	-0,91	1,00	0,26	-0,47	-0,06	-0,06	0,88
TR3	-0,05	0,88	0,90	0,80	0,26	-0,38	-0,63	0,26	1,00	0,42	-0,87	-0,87	0,43
TR4	0,08	0,71	0,70	0,77	-0,06	-0,70	0,20	-0,47	0,42	1,00	-0,56	-0,50	-0,38
PTh	0,18	-0,90	-0,89	-0,90	0,00	0,58	0,42	-0,06	-0,87	-0,56	1,00	0,92	-0,23
StP	0,12	-0,84	-0,85	-0,83	-0,12	0,62	0,42	-0,06	-0,87	-0,50	0,92	1,00	-0,26
Eff	-0,30	0,18	0,21	0,03	0,54	0,58	-0,89	0,88	0,43	-0,38	-0,23	-0,26	1,00

Table 9: Correlation matrix between the objective descriptors

To avoid these multicollinearity issues, we chose to use the PLS regression.

5.2.1 PLS model

The PLS (Partial Least Squares) regression is based on the same principle as the multiple linear regression. Consider a variable Y (for example Softness) and a block of variables X (in our case the objective descriptors), the column of X being the different variables. The PLS regression explains the variable Y by the variables X and is based on the following regression equation:

$$Y = X.\alpha + \varepsilon$$

First, the PLS regression method, described in [11], consists of finding orthogonal component t_h, linear combinations of variables X, that best explains the variable Y and the variables X. Then, the regression equation is obtained by regressing the variable Y on the components t_h and projecting this regression onto the variables X.

More precisely, Y is the subjective variable (column vector with the note for each reed) and X is the table of objective variables (13 columns, one by descriptor, and 20 rows, one by reed). All the variables are presumed to be centered and reduced. After choosing the number of components we want, the PLS components t_h are computed from the residual:

$$X_{h-1} = X - \sum_{k=1}^{h-1} t_k \beta_k^t$$

Where the β_k coefficients are the regression coefficients of X on the components $[t_1, t_2, \dots, t_{h-1}]$ already computed. So we are looking for a normalized vector w_h such that:

$$t_h = X_{h-1}.w_h$$

And maximizing the criterion of covariance:

$$cov(Y, X_{h-1}.w_h)$$

The normalized w_h vector maximizing this criterion is given in [11] by the formula:

$$w_h = \frac{X_{h-1}^t . Y}{\left\| X_{h-1}^t . Y \right\|}$$

After computing the h components, we look for exprime components t_h depending on X ($t_h = X.w_h^*$). The matrix $W_h^* = [w_1^*, \dots, w_h^*]$ is then computed from the matrix $W_h = [w_1, \dots, w_h]$ and $P_h = [p_1, \dots, p_h]$, where $p_h = X^t t_h / t_h' t_h$ with the formula:

$$W_h^* = W_h (P_h^t W_h)^{-1}$$

As we write the regression of y on the t_h components:

$$Y = t_1 c_1^t + \dots + t_h c_h^t + Y_h$$

Where $c_k = Y^t t_k / t_k^t t_k$ (components are orthogonal) and y_h is the residual. Then we can obtain the PLS regression equation depending on X:

$$Y = X[w_1^* c_1^t + \dots + w_h^* c_h^t] + Y_h$$

$$Y = X W_h^* C_h^t + Y_h \quad \text{where} \quad C_h = [c_1, \dots, c_h]$$

$$(5) \quad \boldsymbol{Y = X.\beta + \epsilon} \quad \textit{where} \quad \boldsymbol{\beta = W_h^* C_h^t} \quad \textit{and} \quad \boldsymbol{\epsilon = Y_h}$$

So we obtain a linear equation from orthogonal components, which are linear combination of the objective descriptors. So we have avoided multicollinearity issues.

N.B.: To find the first PLS component, the residual being equal to X, we have to maximize the criterion:

$$cov(Y, X.w_1)$$

Which can be written:

$$cov(y, X.w_1) = corr(y, X.w_1) * \sqrt{var(y)} * \sqrt{var(X.w_1)}$$

And then, for h=1, the PLS method appears as a compromise between a multiple regression of Y onto X (by maximizing $corr(y, X)$) and searching for the first principal component (by maximizing $var(y, X)$). And other components are then sought iteratively, which are orthogonal to the previous ones.

5.2.2 Results of PLS model

In order to choose the optimal number of PLS components and assess the quality of the predictive model, we use a leave-one-out cross-validation and minimize the cross-validation criterion: the PRESS (PREdiction Sum of Squares). The principle is to remove a reed i (a row) from the matrix X and Y and we compute the vector $\beta^{(i)}$ of coefficient of the linear equation for different number of components with the remaining reeds. And we compute the prediction error on the reed i: $E^{(i)} = Y_i - X_i \beta^{(i)}$. The total prediction error is determined by the PRESS:

$$PRESS = \frac{1}{n}\sum_{i=1}^{n} {E^{(i)}}^2$$

We choose the number of components such that the PRESS is minimum. The value of PRESS allows us to compare different predictive models. Lower the PRESS, better the model. Once we have the optimal number of components, we can calculate another indicator of the quality of the validation with the data from the cross-validation. For each reed removed at each step, we compute the coefficient of determination of the cross-validation data ${R_{cv}}^2$:

$${R_{cv}}^2 = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_\imath - \bar{\widehat{Y}}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \text{ with } \widehat{Y}_\imath = X_i\beta^{(i)}, \quad \bar{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_\imath \quad \text{and} \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

The higher the value of ${R_{cv}}^2$, the better the model in cross-validation. In terms of indicators for the performance of the model with all the reeds, we have the fit R² and the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{n}\left(\widehat{Y}_\imath - Y_i\right)^2}$$

We performed a PLS model for the average values of the reeds according the softness, the brightness and the two groups of quality for the subjective descriptors and using the coefficients of the reed effect α_i of the global model (equation (4)), and of two individual models (one per musician) using the equation (3) for the players PK and GS, and the values coming from the computation of the GAMMA method as objective variables. The performance results of the PLS models are given in Table 10. In green are the best values for every indicator and in red the worst.

In general, the individual models (PK and GS) have irregular performances, which is not the case with the global model (Glob ANOVA and Gamma) that have generally good performances. So the global models seem to have smoothed the variability of the global measurements.

More precisely, for the quality models, we can see that the fit for quality models is worse than for the descriptors Softness and Brightness. It seems difficult to create good predictive models for global quality. This may be several reasons for this. It may be that the subjects had very different perceptions of quality (even inside the groups) and that the descriptor is not precise enough. We may have to consider other objective descriptors to make the predictive models. Inside the global quality table, we see that the highest values of ${R_{cv}}^2$ (and the lowest values of PRESS) are obtained by global ANOVA model and Gamma model. So these global models seem to have a better predictive power in simple regression

For Softness, PK model seems to be the best, but we also see that the global models have good performance in cross validation because they have values of ${R_{cv}}^2$ equal to 0.67 and 0.52. The GS model has the worst performance.

For Brightness, for the individual model it is the contrary, PK has the worst performance et GS has as best performance as the global models, hence the irregularity of the individual models. Gamma model has the best performance for this descriptor.

In conclusion, the PLS methods seems to demonstrate good predictive power for the subjective descriptors Softness and Brightness. But we have an issue as for the interpretation of the quality of the model for the validation. As a matter of fact, the PRESS and the $R_{cv}{}^{2}$ are useful tools to compare models between us, but it is not easy to determine the absolute quality of our model in terms of prediction. So we can use the same approach to build a qualitative model which will be easy to interpret. However, the qualitative model won't give better results than the simple PLS regression because it is based on the same data: we transform quantitative variable into a qualitative variable (groups of reeds).

	Softness			
Model	Glob ANOVA	Gamma	GS	PK
Nb component	2	1	1	2
Adjustment R²	0,71	0,62	0,55	0,78
RMSE	0,97	1,09	1,20	0,84
PRESS cv	1,39	1,66	1,84	1,05
R² cv	0,67	0,52	0,45	0,79

	Brightness			
Model	Glob ANOVA	Gamma	GS	PK
Nb components	1	1	1	1
Adjustment R²	0,67	0,68	0,65	0,59
RMSE	0,65	0,63	0,67	0,73
PRESS cv	0,54	0,49	0,49	0,77
R² cv	0,68	0,71	0,69	0,57

	Quality gp1			
Model	Glob ANOVA	Gamma	GS	PK
Nb component	1	1	1	1
Adjustment R²	0,52	0,58	0,54	0,45
RMSE	0,69	0,64	0,67	0,74
PRESS cv	0,63	0,5476	0,76	1,17
R² cv	0,47	0,56	0,41	0,13

	Quality gp2			
Model	Glob ANOVA	Gamma	GS	PK
Nb component	1	1	1	1
Adjustment R²	0,36	0,35	0,26	0,41
RMSE	1,19	1,19	1,26	1,13
PRESS cv	1,71	1,66	1,82	1,87
R² cv	0,30	0,31	0,21	0,31

Table 10: Performance results for the PLS predictive models

5.3 Qualitative model

If the PLS approach was originally used in regression models, it can be also used as a classification method [12]. So we decided to create a qualitative model by making classes of reeds according to the descriptors "Softness" and "Brightness". This model will be easier to assess in terms of performance because we will obtain a percentage of well classified reeds which we can more easily interpret.

5.3.1 Partitioning of the reeds

First, we have to divide the reeds into several classes according to the descriptors Softness and Brightness. To achieve this, we applied Hierarchical Ascendant Clustering on the subjective notes of Softness and Brightness to the row data, using a Euclidian distance and the Ward criterion. We obtained the following classification presented in Figure 19. We see that we can separate the reeds into three classes. What mainly characterizes the three classes is Softness (we obtain the same classes if we take only Softness as subjective descriptor). The class at the left is the class of soft reeds (the majority), the class at the right is the class of hard reeds, and the class at the middle is the class of "middle" reeds.

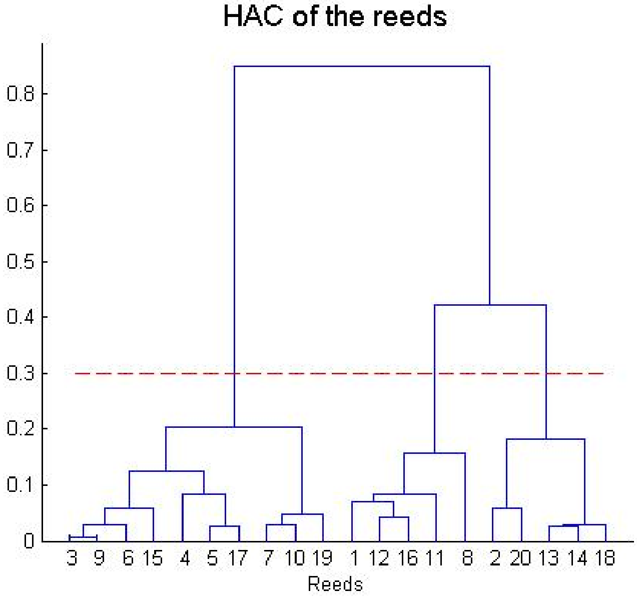


Figure 19: HAC of the reeds according to the descriptors softness and brightness.

In the Figure 20 are presented the plan of reeds according to the brightness and the softness with the classes circled.

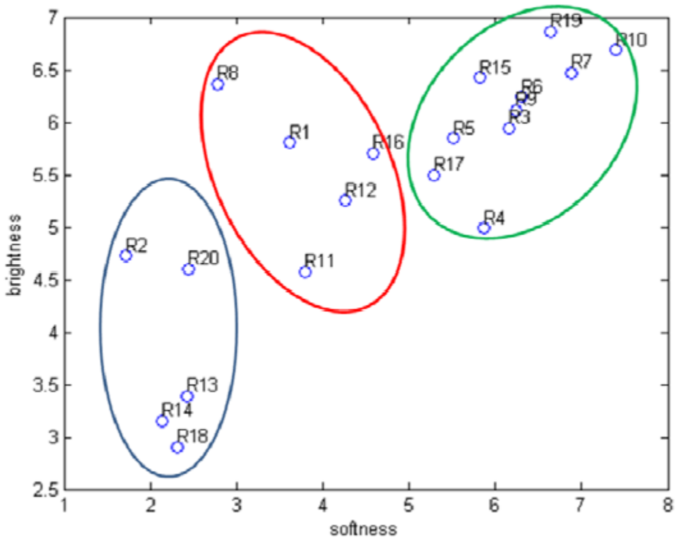


Figure 20: Subjective classes of the reeds

5.3.2 Principle of PLSDA

We performed a PLSDA (Partial Least Square Discriminant Analysis) method on these 3 classes. The purpose of this method is to find the PLS components that best explain the different classes. It can be linked with Linear Discriminant Analysis. The approach is exactly the same as for a PLS regression, except that the variable Y (see part 5.2.1) is not the vector with the quantitative value of a subjective descriptor anymore but an array with as many columns as classes and as many rows as reeds (coding of the qualitative group-disjunctive form). This array Y in each column takes the value 1 if the reed (row number) is in the class and 0 if it is not. After that, it is exactly the same algorithm employed as for the PLS regression. With the components we computed, we estimate $\hat{Y}$ and for each row (each reed), the column of the largest component will correspond to the class of the reed. Here the strategy employed is the most common employed with the PLSDA method: the Max Indictors strategy [13].

But this method often overfits the data and we have to be careful about the validation. According to [14], the most accurate method of validation is to perform a cross-validation on a training set (in order to select the optimal number of PLS components), and after make a validation on a test set composed of individuals who were not in the training set and ho were not used to computed the model. But we have only 20 reeds, that is not enough to constitute a training set and a test set. So we employed a method of simple cross-validation as before to choose the optimal number of PLS components and to assess the quality of our classification. We used the “leave-one-out” method but we changed of indicator of quality. For a classification task we cannot use the PRESS as previous. Instead that, we remove one of the 20 reed, we perform the PLSDA on the remaining 19 reeds, and we use the obtained model to classify the reed removed and we see if it is well classed. We repeat the operation for the 20 reeds and we obtain a percentage of well-classified reeds.

To assess the quality of the model we obtain with respect to random data, we also randomized the reeds labels. The principle is to take the array Y and to perform a random permutation of the rows. After that the classification is performed. So we make a classification on random data and we look at the percentage of well-classified reeds for this random classification of the reeds. The result obtained in the tables is the mean over 10 randomizations. We do this randomization for the entire model with all the reeds and in the cross-validation part. If we have a low percentage of well-classified reeds in this randomization test, we can say that our first classification makes sense because our model will have done better with the real data than with random data.

5.3.3 Results of the models

For the objective measures, we used exactly the same data as for the regression analysis. The results of the PLSDA classification are given in Table 11. The row "Nb component" corresponds to the optimal number of PLS a component from the cross-validation, the row "% well-classified" corresponds to the percentage of well-classified reeds using all the reeds for the classification (no validation set). The "% with randomization" is the mean of well-classified reeds over the 10 randomizations (again all the reeds are used). The "% well classified cv" row is the percentage of well-classified reeds using the leave-one-out cross-validation. And finally, the "% with randomization cv" row is the mean of well-classified reeds over the 10 randomization in the cross-validation part.

	PLS DA			
Model	Glob ANOVA	Gamma	GS	PK
Nb component	4	6	6	4
% well classified	85	90	80	90
% with randomization	41	43	43	41,5
% well classified cv	80	75	60	80
% with randomization cv	40	25	35	45

Table 11: Performance results for the PLSDA model.

We see that we have the best results for the global models and the individual model of PK. As a matter of fact, in cross validation, the percentage of well-classified reeds is 80% for the Global ANOVA model and for the PK model which represents 16 of the 20 reeds. The Gamma model has also an important predictive power with a percentage of 75% in cross-validation. The GS model gives worse results. We see that all our predictive models give much better results than the model with the randomizations, which has a percentage of well-classified reeds of around 40% using all the reeds and a lower percentage for using the randomization in the cross-validation. This are low percentages. So our predictive models make sense and the validation process is relevant.

Let's see now the variables that have an important contribution in the classification using the matrix β of the PLS coefficients (equation (5)). In Table 12 is presented this matrix β using the reed coefficients α_i global ANOVA model as objective variables. There is one column per class and one objective variable per row, so we can see which objective variables best explain each class. In green are the three highest values for each class. Recall that we had three classes: class 1 contains the softest reeds, class 2 the intermediate reeds and class 3 the hardest reeds. So we will call these three classes the soft class, the medium class and the hard class, respectively. It is the names employed in the columns of Table 12 .

		Soft Class	Medium Class	Hard Class
Descriptors	AtT	-0,058	0,032	0,026
	SC	-0,517	0,599	-0,082
	OSC	-0,575	0,612	-0,037
	ESC	-0,076	0,218	-0,142
	OER	-0,222	0,262	-0,040
	Lv	-0,335	1,079	-0,744
	TR1	0,103	-0,162	0,059
	TR2	-0,059	0,123	-0,064
	TR3	-0,043	0,039	0,004
	TR4	-0,001	0,000	0,001
	PTh	-0,376	0,259	0,117
	StP	0,167	-0,198	0,030
	Eff	-0,008	0,030	-0,023

Table 12: Matrix of the PLS coefficients of the predictive model using the values of the global ANOVA model

The objective variables that best explain the soft class are Spectral Centroid, Odd Spectral Centroid and Pressure Threshold. These were the variables having the highest classic correlations with the subjective descriptor (Table 7), but it is surprising to note that the coefficients for SC and OSC are negative. For the medium class, the main objective variables are SC, OSC and the Spectral Amplitude Lv. And finally, for the hard class, the main variables are Even Spectral Centroïd (ESC), LV and Pressure Threshold (PTh). As the PTh have a high negative correlation with the Softness, it seems logical that this objective variable has a good contribution in the explanation of the class soft and the class hard with the sign we have. But the signs of some coefficients as for the SC or the OSC show that the PLSDA method don't always give logical predictive models even though the performance of this model is good as shown as before. We have to be careful if we want to use this model onto other reeds to validate the model.

In conclusion we see that with our global model we can achieve a good predictive model using the PLSDA.

6 Conclusion

In conclusion, a subjective analysis was performed on 20 saxophone reeds. Three descriptors were assessed: Softness, Brightness and Global Quality. Then objective measurements were performed and 13 objective variables were extracted from these measurements. Finally, a predictive model was built using the objective variables to predict the subjective class of the reeds. The results show that our model has a good power of prediction.

In this study, the objective variables were chosen from the previous studies performed on the saxophone reeds[3][1] and we saw variability in the computation of these variables. A future work could consist in finding other objective variables that have not such variability. We also can consider the different musicians who performed the objective measurements as different measurement systems and consider the values of the objective variables new different variables for each musician. We will multiply the number of objective variables by the number of subjects. The number of variables would be important but the PLSDA is a classification method who can deal with a high number of variables.

To conclude, this study can find new orientations to improve the actual predictive model.

7 References

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

[2] 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.

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

[4] 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.

[5] M. Barthet, P. Guillemain, R. Kronland-Martinet and S. Ystad, "From calrinet control to timbre perception," *Acta acustica,* vol. 96, pp. 678-689, 2010.

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

[7] D. Hirst and T. Naes, "A graphical technique for assessing differences among a set of rankings," *Journal of Chemiometrics,* vol. 8, pp. 81-93, 1994.

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

[9] S. Ledauphin, M. Hanafi and E. Qannari, "Assessment of the agreement among the subjects in fixed vocabulary profiling.," *Food Quality and Preference,* vol. 17, pp. 277-280, 2006.

[10] S. Vancolen, "La régression PLS," Master Thesis, Université de Neuchâtel, 2004.

[11] M. Tehenhaus, "L'approche PLS," *Revue de statistique appliquée,* vol. 47, no. 2, pp. 5-40, 1999.

[12] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of Chemometrics,* vol. 17, pp. 166-173, 2003.

[13] S. Chevalier, D. Bertrand, A. Kohler and P. Courcoux, "Aplication of PLS-DA in multivariate image analysis," *Journal of Chemometrics,* vol. 20, pp. 221-229, 2006.

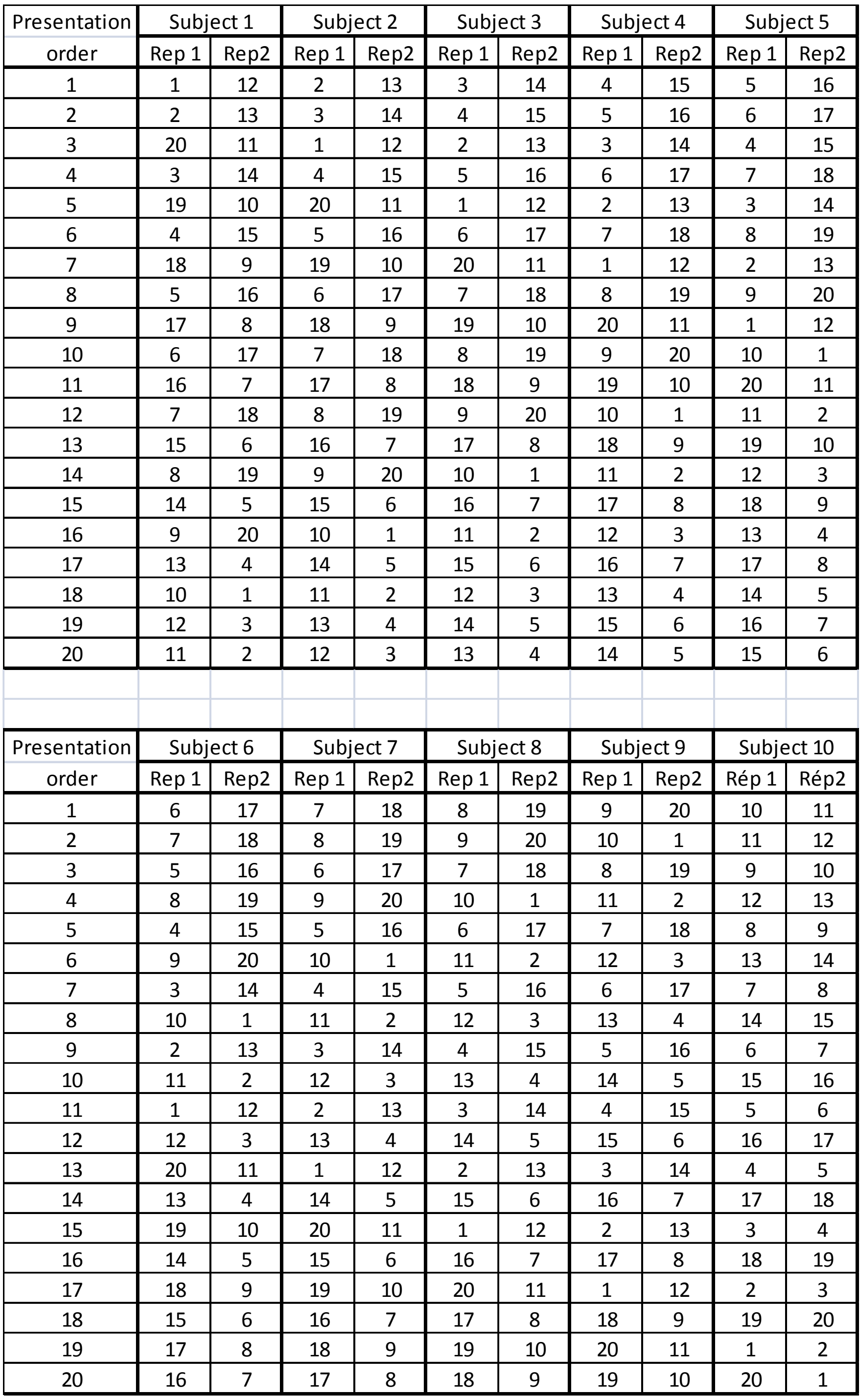
[14] J. Westerjuis, H. Hoefsloot, S. Smit, D. Vis, A. Smilde, E. Velzen, J. Duijnhoven and F. Dorsten, "Assessment of PLSDA cross validation," *Metabolomics,* vol. 4, pp. 81-89, 2008.

[15] G. Dijksterhuis and J. Gower, "The interpretation of Generalised Procrustes Analysis and allied methods.," *Food Quality and Preference,* vol. 3, no. 2, pp. 67-87, 1991/1992.

[16] J. Gower, "Generalized Procrustes Analysis," *Psychometrika,* vol. 50, pp. 33-51, 1975.

Appendix A: Presentation plan of the reeds for the subjective tests

The presentation plan of the reeds followed a Williams Latin Square, it is presented in the Figure 21. We had 10 subject, 2 repetition of assessment and 20 reeds. So we choose to take a Latin Square of size 20. SO we consider 2 different orders for the two repetitions. Consequently, the presentation plan was perfectly balanced.



Presentation	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5	
order	Rep 1	Rep2	Rep 1	Rep2	Rep 1	Rep2	Rep 1	Rep2	Rep 1	Rep2
1	1	12	2	13	3	14	4	15	5	16
2	2	13	3	14	4	15	5	16	6	17
3	20	11	1	12	2	13	3	14	4	15
4	3	14	4	15	5	16	6	17	7	18
5	19	10	20	11	1	12	2	13	3	14
6	4	15	5	16	6	17	7	18	8	19
7	18	9	19	10	20	11	1	12	2	13
8	5	16	6	17	7	18	8	19	9	20
9	17	8	18	9	19	10	20	11	1	12
10	6	17	7	18	8	19	9	20	10	1
11	16	7	17	8	18	9	19	10	20	11
12	7	18	8	19	9	20	10	1	11	2
13	15	6	16	7	17	8	18	9	19	10
14	8	19	9	20	10	1	11	2	12	3
15	14	5	15	6	16	7	17	8	18	9
16	9	20	10	1	11	2	12	3	13	4
17	13	4	14	5	15	6	16	7	17	8
18	10	1	11	2	12	3	13	4	14	5
19	12	3	13	4	14	5	15	6	16	7
20	11	2	12	3	13	4	14	5	15	6

Presentation	Subject 6		Subject 7		Subject 8		Subject 9		Subject 10	
order	Rep 1	Rep2	Rep 1	Rep2	Rep 1	Rep2	Rep 1	Rep2	Rép 1	Rép2
1	6	17	7	18	8	19	9	20	10	11
2	7	18	8	19	9	20	10	1	11	12
3	5	16	6	17	7	18	8	19	9	10
4	8	19	9	20	10	1	11	2	12	13
5	4	15	5	16	6	17	7	18	8	9
6	9	20	10	1	11	2	12	3	13	14
7	3	14	4	15	5	16	6	17	7	8
8	10	1	11	2	12	3	13	4	14	15
9	2	13	3	14	4	15	5	16	6	7
10	11	2	12	3	13	4	14	5	15	16
11	1	12	2	13	3	14	4	15	5	6
12	12	3	13	4	14	5	15	6	16	17
13	20	11	1	12	2	13	3	14	4	5
14	13	4	14	5	15	6	16	7	17	18
15	19	10	20	11	1	12	2	13	3	4
16	14	5	15	6	16	7	17	8	18	19
17	18	9	19	10	20	11	1	12	2	3
18	15	6	16	7	17	8	18	9	19	20
19	17	8	18	9	19	10	20	11	1	2
20	16	7	17	8	18	9	19	10	20	1

Figure 21:Plan presentation of the reeds for the subjective tests.

Appendix B: Score of pattern presented for the subjective tests

In the Figure 22 is presented the score of pattern presented to the subject during the subjective tests. For the softness, the pattern was composed of notes in the low register at the dynamic piano. As a matter of fact, it is difficult to produce a sound in the low register at this dynamic so it is easier to see how hard a reed is.

For the brightness, the pattern was composed of en arpeggio beginning by notes in the high register at the dynamic mezzo forte. It seems that the differences between the reeds were more seeable in this register at this dynamic.

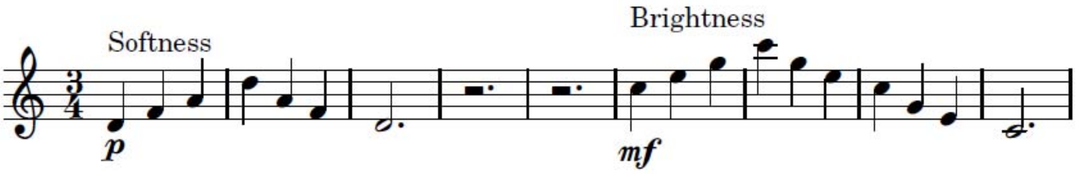


Figure 22: Score presented for the subjective tests in Bb

Appendix C: Score of notes played during the objective measurements

In the Figure 23 is presented the score of notes played during the objective measurements.

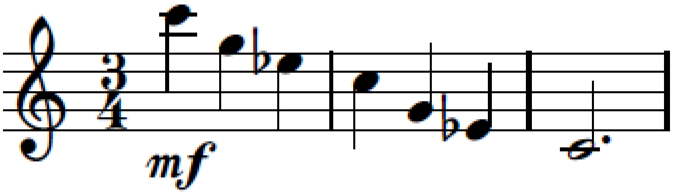


Figure 23: Notes played during the objective measurements (concert key).

Appendix D: The GAMMA method

Method

Instead of making a raw average value of the assessments, the average value can take into account weights for each subject, according to their agreement with the panel[9]. It is based on a fixed vocabulary, common for all the assessors. The interpretation of the common configuration is in this way easy, because it corresponds directly to the descriptors used by the subjects. It consists in computing weights for each subject, which are representative of the degree of agreement of the subject with the rest of the panel.

Let X_i denote the (I×M) matrix, describing the assessments made by assessor i on the I products according to M descriptors. Matrix X_i is called a configuration. The first step of the method is, for each configuration, to center (removing the effect of the judge on the scoring) and rescale (standardize the data to the same total variance) the data.

- center the data : substract, for each data, the average of the column of the X_i matrix. This will give matrix Xc_i.
- standardization: multiply all the data by the factor $\theta_i = \frac{1}{\sqrt{trace(Xc_i^t.Xc_i)}}$. This will give matrix Y_i.

To assess the similarity between two subjects i and j, the method computes the following quantity that is directly the correlation coefficient between Y_i and Y_j, considered as vectors by rearranging the data in a single column.

$$t_{ij} = \frac{trace(Y_i^t.Y_j)}{\sqrt{trace(Y_i^t.Y_i)}.\sqrt{trace(Y_j^t.Y_j)}}$$

The similarity matrix S_{ij} of size (J×J) between all the assessors if given by:

$$S_{ij} = \frac{1+t_{ij}}{2}$$

The eigenvector $\beta = [\beta_1, \beta_2, \dots, \beta_J]^t$ associated with the largest value, represents the degree of agreement of the subjects with the rest of the panel. A weighted average configurations is finally computed to characterize the products:

$$C = \sum_{j=1}^{J} \beta_j . Y_j$$

C is next subjected to a PCA to characterize the differences between products.

To characterize the performance of the panel, two indices are computed,

$$\alpha_i = \frac{trace(Y_i^t.C)}{\sqrt{trace(C^t.C)}}$$,the correlation coefficient between Yi and C

$$\gamma = \frac{1}{J}\sum_{j=1}^{J}\alpha_j$$: the average correlation coefficient for the panel (performance index)

Results

With the whole group of 10 subjects, the values of the alpha coefficient for the average value across the sessions are given in Figure 24.

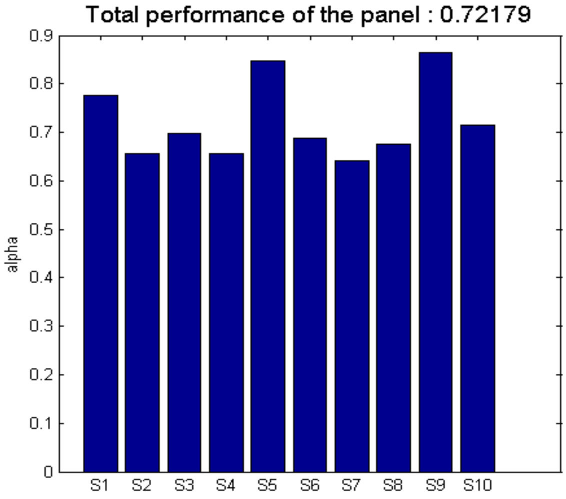


Figure 24: plot of the alpha coefficient for the whole group of 10 subjects

The performance index of the panel is 0.72, which is good. We can say we have a good panel despite the difficulties of reliability with some subject with the brightness. It is not surprising to see that S5 and S9 are the subjects with the highest alpha, they were among the most reliable subjects (see section 3.3.4).

To characterize the products, the matrix C_{group2} of the weighted configurations is computed, and a representation of the matrix C_{group2} is given in Figure 25.

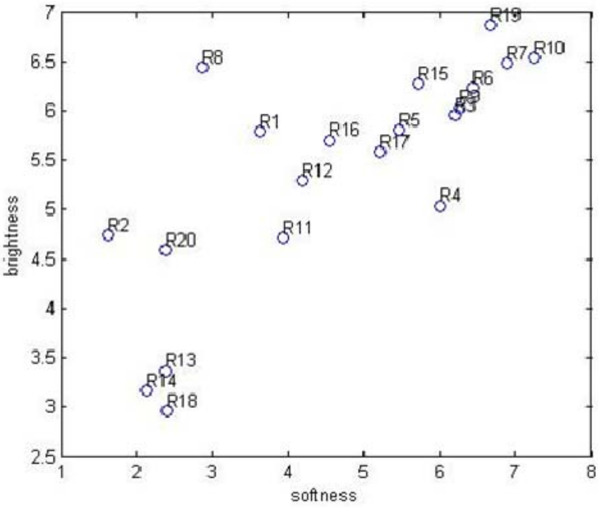


Figure 25: Plan of the reeds according to the softness and the brightness (gamma method)

There is clearly a positive correlation between the two dimensions, but one dimension takes 88% of variance: two dimensions are necessary to represent in particular the atypical behavior of reed 8 (not soft, and very bright).

- R10 , R7, R19 are the most soft and bright reeds
- R14, R18, R13 are the less soft and bright reeds

These results are in accordance with the average value. Finally, the GAMMA method does not bring a different result, because the agreement between the subjects is high.

Appendix E: Generalized Procrustes Analysis

GPA is a multivariate technique for the analysis of free-choice profiling data (FCP[2]), but it has also been applied to consensus vocabulary profiles (our case for the saxophone reeds). The aim is to study the consensus among experts, to assess scale use, attribute interpretation, panel performance, monitoring…

It also allows one's to compare the proximity between the terms that are used by different experts to describe products. The GPA method was first described by [15], interpretation of GPA can be found in [16]. Let $X_i$denote the (I×M) matrix, describing the assessments made by assessor i on the I products according to M descriptors. Note that with the Free Choice Profiling (FCP), the variables which describe the products are not necessarily the same, the number of variables can also be different for each configuration. GPA is a method for producing a consensus configuration $\bar{X}$ from the set of *J* different individual data matrices, and to represent the consensus via PCA. The principle of GPA is to apply transformations (translation, isotropic scaling, rotation/reflection) to the configurations X_i so as to minimize a goodness of fit criterion (the distance between the transformed configurations X_i and the consensus configuration $\bar{X}$). GPA only allows 'rigid-body' transformations to the datasets and respects the relative distances between products. The individual and consensus configurations are typically submitted to PCA and projected onto a lower dimensional space. This space provides a vantage point to compare individual data and to visualize the consensus.

The degree of consensus is assessed by studying the variance of the datasets. The total variance V_T can be partitioned as follows (equation (9)):

$$(6)\quad \boldsymbol{V_T = V_C + V_W + V_R}$$

Where V_C denotes the variance of the consensus, V_W the within-product variance in the projection space an V_R the residual variance. By dividing by V_T, and sharing the within variance V_W among the I products, the equation becomes (equation (10)):

$$(7)\quad \boldsymbol{100\% = R_c + \sum_{j=1}^{I} r_{jW} + R_R}$$

R_c corresponds to the consensus ratio: a large R_c indicates a good consensus.

$\boldsymbol{r_{jW}}$ indicates the within variance of product j. A small $\boldsymbol{r_{jW}}$ indicates a bad consensus for this particular product j.

By considering the configurations of the 10 subjects, a GPA gives a consensus ratio of 55,5%. This variance ratio is significant with the permutation test.

The plane of the first two factors of PCA (consensus plane) is given Figure 26. The results are in accordance with the average configuration and the gamma method:

- R10 , R7, R19 are the most soft and bright reeds
- R14, R18, R13 are the less soft and bright reeds

[2] FCP: Under this type of sensory profiling, each assessor or judge describes a product's characteristics using his/her own list of sensory attributes

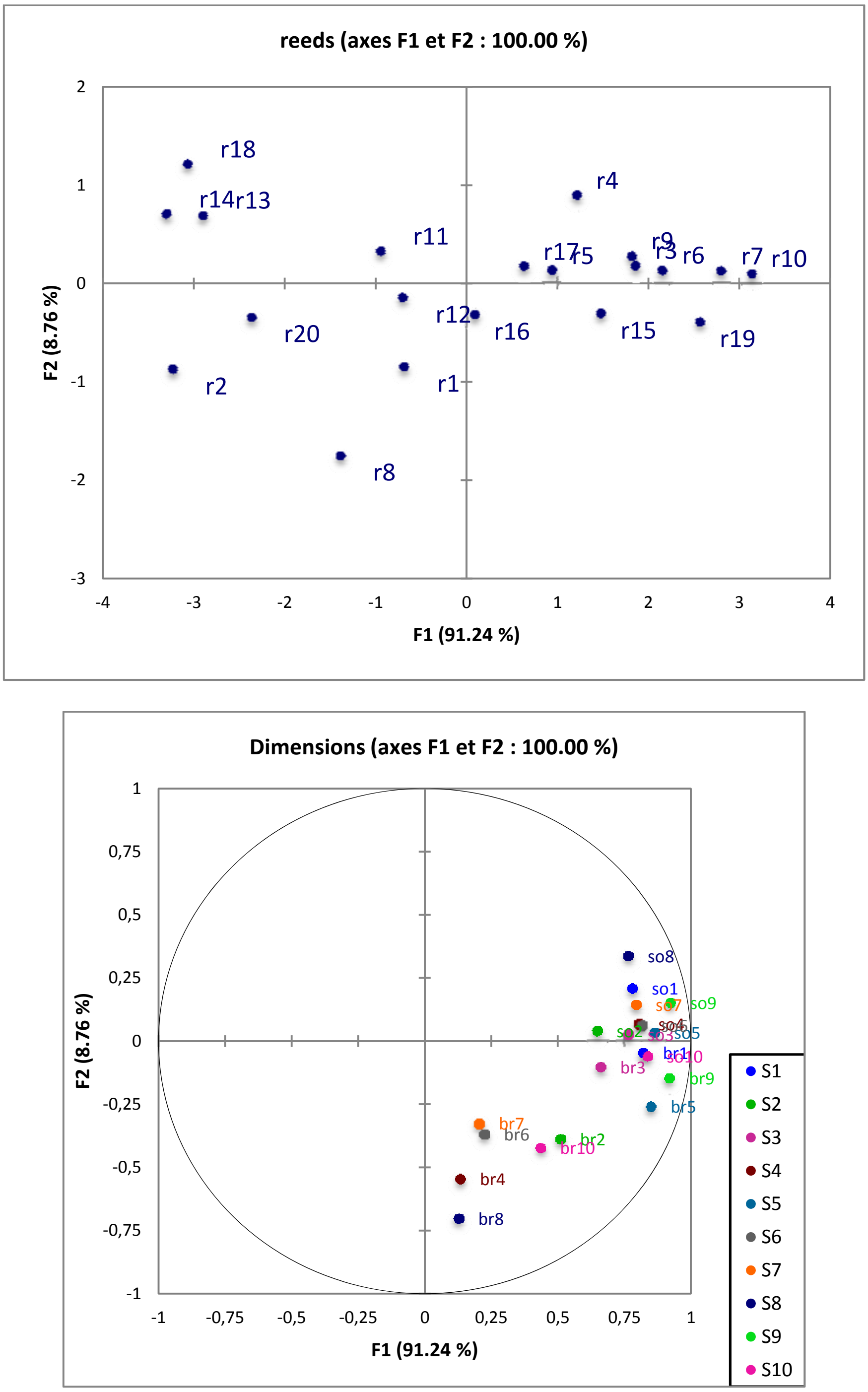


Figure 26: PCA of the consensus after GPA.

Factor 1 can be interpreted as softness, and factor 2 as brightness.

Finally, GPA does not bring different conclusions than the GAMMA method or the simple average.