# Assessment of Timbre Using Verbal Attributes

Graham Darke, Faculty of Computing, Engineering and Mathematical Sciences,
University of the West of England, Bristol, UK
Graham2.Darke@uwe.ac.uk

## Abstract

As part of the perceptual processing to form an abstract picture of a sound, the question arises as to how various timbral judgments are made. An early step in these 'processing stages' suggests that comparisons are made with our previous experiences of sounds. These experiences are often given descriptive labels by which we can communicate the concept of a timbre. It follows that a listener might be able to make qualitative assessments of a sound when asked to judge against the perception of this set of verbal attributes.

This study looks at the effectiveness of communicating assessment judgments concerning timbre under controlled conditions. The work presented reports on a study where a group of musicians were asked to assess a number of musical sounds against 12 verbal attributes. The study is based around describing and analysing sounds from a subset of pitched European orchestral instruments.

Using statistical analysis, the collated judgements are analysed to indicate the level of agreement that can be obtained from a collective view of each attribute. Some evidence of agreement between assessments is shown by the use of group means and variances, along with a preliminary discussion on a few of the reasons behind the inconsistencies.

## Introduction

Much research concerning timbre perception has been oriented towards explaining the dimensions of a multidimensional scaling (MDS) map of timbre similarity ratings (e.g. Grey and Gordon, 1978). The final analysis of these MDS dimensions has largely been by interpretation, often based on comparison with spectral analysis. This interpretation is not always intuitive towards a verbal description and has led to a debate on how many dimensions are relevant to timbre. The debate centres around the selection of the attribute rating scales that provide greatest differentiation between timbres and the ability to distinguish different levels of each rating scales within the experimenter's sound data set. It ignores a large part of a timbre's constitution that is important only in differentiating between other sounds. Synthesis of new sounds requires more detail if construction from the basic timbral components is to make these sounds indistinguishable from the original specification.

The aim of this paper is to study the effectiveness of communicating assessment judgments concerning timbre under controlled conditions; is there a common understanding of the way we perceive and describe timbre. The paper reports on an experiment where a group of musicians, both amateur and professional, were asked to assess a number of musical sounds (one single note at a time) against 12 verbal attributes. The musicians who participated in this study used UK English as their primary spoken language. The study is based around describing and analysing sounds from a subset of pitched European orchestral instruments.

Given sufficient confidence in the data obtained, it is anticipated that correlation with the time and frequency representations of the acoustic representations will provide a set of 'attribute predictors'. These predictors can then be compared with listener evaluation for new sounds as an iterative validation of the process.

## The Psychology of the Perception of Sound

An information processing account of perception seeks to delineate the stages of processing through which a stimulus passes and determine the properties of the representation at each stage. This indirect link between the physical sound itself and our comprehension of the sound implies that our understanding of sound (and hence timbre) is a subjective, abstract concept.

A number of cues are used for the perception and identification of sound, dependent on the context. By monitoring this context change under different circumstances, M. Pitt has shown that acoustic information is encoded perceptually at the central level and that this central 'processor' adapts to sound input (Pitt, 1995). This work helps explain the representation of related sounds in auditory perception. However, these representations do not necessarily correlate with linguistic description processes.

As part of the perceptual processing to form this abstract picture of a sound, the question arises as to how various timbral judgments are made. An early step in these 'processing stages' suggests that comparisons are made with our previous experiences by comparison with other sounds, situations or impressions. These experiences are often given descriptive labels by which we can communicate the concept of a timbre. It follows that a listener might be able to make qualitative assessments of a sound's impression when asked to judge against the perception of this set of verbal attributes.

## Implications for Musical Practice

Given suitable verbal descriptors, quantitative analysis allows parts of a sound's timbre to be rated and compared by different observers. The resultant analysis gives a framework by which timbre can be described using a common 'verbal descriptive' language. Such a language would be highly valuable in describing composition and performance, synthesis control and studio technology.

This research is a study of the attitude of a group of English-speaking musical participants on a subset of the available musical sounds. It is mainly qualitative in approach, since it is an attempt to better characterise the attributes of timbre. Later work is expected to extend the range of cultural and linguistic backgrounds and to extend our understanding of other possible sounds.

## Implications for Musicological Interdisciplinarity

We are perpetually confronted by new sounds that we attempt to describe with reference to what we have already heard. A method of more fully describing the timbre of these sounds, and their relationship to acoustical form, would provide a common point of reference for different disciplines. Thus, timbral descriptors would enable composers, performers, musical instrument manufacturers and scientists to communicate effectively the nature of different timbres. Gaining a better understanding of the relationship between acoustical sound and its perception is also fundamental to further timbral research study. This paper studies the problem of understanding timbral descriptors by asking the question '*Can we ask a group of people to describe a set of musical sounds and arrive at a common understanding?*'

## The Nature of Timbral Description

The traditional view of timbre is in the perception that two sounds are recognised as different other than by pitch and other than by loudness. In order to communicate these differences, the first step is to define a vocabulary. There are three basic types of verbal descriptors of timbre:

### Sound itself (onomatopoeia)

This group describes the sounds as one hears it. Animal sounds typically fall into this category, as does the 'Wha-Wha' effect of certain muted jazz trumpet sounds.

Comparison of a timbre to real instruments (violin-like or flute-ish) is another method often used. These descriptions lead to pair wise comparison tests, but with the drawback that there is no one definition of what a violin or a flute should sound like.

### Sounding situation (sound source)

This group describes sounding situations including sound sources. They can be classified into the sub-groups:

- **What** (blown wind instruments)
- **Where** (on a barge on the Thames)
- **When** (to the King)
- **How** (unmuted)
- **Event** (regal music for the Royal Party)

Although this group paints in words a picture of the sound that is perceived, the picture provided is not easily generalisable in the timbre domain.

## Sound impression (adjective)

This group describes one's subjectivity. Three sub-classifications are again available here:

- **Adjective / adverb** (e.g. brassy)
- **Subjective word** (e.g. "baroque sounding")
- **Figurative descriptions** (e.g. "On parade again")

Sanae Wake and Toshiyuki Asahi have assessed these word groupings for use in a sound retrieval system. Sound retrieval has commonality here to timbre identification – that of communicating in an effective manner the position of a sound in timbre space. (Wake and Asahi, 1998)

Wake and Asahi found that 86% of descriptions of a sound belonged to the 'sounding situation' group of words. However, the scope of the current study is to look at descriptive words that are deemed generalisable across the timbre domain. Thus, words that belong to the adjective group of sound impressions have been selected.

## Sound Impression – Verbal Attributes

It has already been discussed that timbre has multidimensional qualities that may be partially described using adjectives. The postulate here is based on the measurement of meaning by Osgood that:

> this semantic space is a region of some unknown dimensionality, and is Euclidean in character. Each semantic scale, defined by a pair of polar (opposite in meaning) adjectives, is assumed to represent a straight-line function that passes through the origin of this space
> (Osgood et al, 1957, p. 25)

Although semantic space is being applied to test the validity of a timbral description, the Euclidean character and straight-line function through the semantic origin for each scale have yet to be proven for this study. In other words, this semantic space is the projection of an adjectival (or adverbal) schema onto a specified set of sounds. A common verbal rating technique is to utilise a scale of extremes of which an adjective and its purported opposite are placed. This technique is known as the **semantic differential**.

Having initially studied semantic differentials, it occurred to Kendall and Carterette that the purported opposite of an attribute is not necessarily its antipode (Kendall and Carterette, 1993). What is the opposite of bright? Dim, dull or muted? Their conclusion was that, although not specifically a methodological study, purported opposites confused the understanding of the timbre picture. The constraint applied by Osgood requires that a measurement scale pass through the origin. The use of direct opposites (e.g. '*Bright <--> Not Bright*') provided greater differentiation between individual timbres. The estimation of timbre using these direct opposites is called 'Verbal Attribute Magnitude Estimation' (VAME).

## Listening Experiments

Durand Begault defined listening experiments as:

> "An attempt to define a cause-effect relationship between "physical" (design engineering) parameters of auditory stimuli and subjective responses"
> (Begault, 2003)

The evaluation of these subjective responses, as described by Begault (Begault, 2003) can take several forms:

1. Discriminative; Are sounds different in any way?

   Two alternative forced choice, paired comparison tests. For example, which sound sample is most like an oboe? How much do two sounds differ? Many researchers have used this branch of listening tests as the input to multi-dimensional scaling (MDS) to ask the question 'How do sounds relate to each other?'

2. Descriptive; how do sound examples differ in specific sensory characteristics?

   This group of listening test methods (psychometrics) covers the scaling (rating of specific attributes) and magnitude estimation requirements of 'VAME'. The subject is asked to quantify the sensory magnitude of a stimulus on a one-dimensional auditory attribute.

3. Affective; how well is sound stimulus liked; which of two stimuli is preferred?

   The attitude testing method here is by scaling (rating of overall quality on a 7 point scale) & paired comparison.

   Method of preference quantification - the subject task is to rank order the stimuli in order of preference. Rather than asking the subject to describe a sound, they are asked to give an opinion or attitude towards that sound. Loudspeaker preference tests are covered in this group as a result of Likert-style analysis.

Since Verbal Attribute Magnitude Estimation is a form of descriptive test and the response is subjective, there is an implied quantification of the stimuli. Whereas pitch is measured in Hertz and loudness in phons, timbral attributes do not have any defined scale of measurement.

## Experiment Reproducibility

No measuring instrument is ideal – whether it be a human subject or otherwise. In the case of psychometrics, the experimenter has an additional variance; the uncertainty of measuring the meaning of an attribute under test. When a subject is testing a sound for loudness, how is the paradox between a loud timbre and a loud sound resolved? According to Kendall and Carterette, the attribute '*loud*' appeared to be associated with '*hard*' and '*heavy*' attributes, even though the synthetic inputs had all been equalised for loudness (Kendall and Carterette, 1993).

In the measurement of meaning, Osgood (Osgood et al, 1957, p. 125) identified the following evaluation criteria:

- **Objectivity**

How brassy is a tuba if one knows how it is constructed?  What about its sound?

A method is objective to the extent that the operations of measurement and means of arriving at conclusions can be made explicit and hence reproducible.

It may be argued that the data obtained from semantic measurements are essentially subjective - and all that has been done is to objectify expressions of these subjective states. This is true, but it is not a criticism of the method. Objectivity concerns the role of the observer, not the observed.

- **Reliability**

Reliability measures the degree to which assessment scores can be reproduced when the same objects are measured repeatedly. In a test – retest scenario, does a subject respond reliably? Is a subject affected by a halo effect from previous judgements or is each assessment independently and intuitively made?

Does each subject rate an attribute equally? While some attributes have defined reference points, other attributes do not. Loudness is defined as 0 phons (barely audible threshold) and 120 phons (pain threshold). '*Bright*' and '*not bright*' attributes are not so defined; how does a score of 5 obtained from one subject correlate with a score of 4 by another for the same sound?

- **Validity**

An assessment is said to be valid when it measures what it is supposed to measure. A more refined and quantitative statement is that an assessment is valid to the extent that scores on it correlate with scores on some criterion of that which is supposed to be measured. Loudness (intensity) has already been mentioned as including a component of timbral heaviness as well as intensity.

Ideally, "Verbal Attribute Magnitude Estimation" (VAME) scores should be correlated with some independent criterion of meaning - but there is no commonly accepted quantitative criterion of

meaning. In lieu of such criterion, initial studies for each attribute can only be assessed on "face validity".

- **Sensitivity**

An assessment is sensitive to the degree that it can discriminate between adjacent sounds. Discriminative observation tests use 'Just Noticeable Differences' as the smallest unit of measure in dissimilarity ratings. In magnitude estimation, the degree to which verbal attributes can be differentiated is not known; as in pitch discrimination, sensitivity to 'just noticeable differences' may be greater than when estimating absolute magnitude standalone.

Sensitivity implies both reliability and validity.

- **Comparability**

This is again an extension of the notion of validity - over how broad a range of situations are the measured assessments equally valid. In the case of VAME, the range of situations is mainly 'subjects' and 'concepts'. Can the magnitude estimates be applied with equal validity and hence comparability.

## The Experiment

The objective of this experiment is to study the underlying evaluation processes for different people during the assessment of timbre:

> *The evaluative system comprises a process of automatic activation of evaluations that is triggered by the mere presence of an object in one's field of perception. A major tenet is therefore that the process of evaluation activation is preconscious. A second postulate is that the evaluative system is functionally dissociable from the perceptual and motivational system.*
> (Musch & Klauer, 2003, p. 7-8)

The underlying evaluation system in the context of this study is the estimation of verbal attribute magnitudes within specific sounds:

- How consistent are the assessments between subjects? Is there correlation between the assessments for each of the verbal attribute – i.e. how similar is the evaluative system between different people – is it learnt by experience or implicit?

- Is each assessment intuitive and automatic? Can a subject arrive at consistent attribute ratings on hearing a sound once?

- Is there any pre-conditioning where the recently heard sounds affect subsequent assessments?

### The Test Procedure

Twenty-two subjects participated in the experiment (ages ranging from 18 – 82), arranged in two groups A and B. Fifteen sounds (drawn from the pitched western orchestral instruments set) were tested against 12 verbal 'differential adjective' style attributes (categories). Since this study is concerned with studying a person's ability to analyse subjectively the quality of a sound, the choice of attributes to be used is, to a certain degree arbitrary. The final choices were based around selecting the attributes that the experimenter felt best described the steady-state 'sound colour' part of a sound:

> How Bright, How Harsh, Has Clarity, How Thin, How Hard, How Full, How Nasal,
> How Muted, How Reedy, How Brassy, How Metallic, How Wooden

Each of the selected sounds was approximately corrected to be of the same pitch (E4 – intonation permitting) and loudness. The sequence of sounds was randomly ordered in each of the categories. All the members of group A assessed the sounds in the same random order. Group B members heard the same sounds but in the reverse order to group A.

Each subject used the same set of sample sounds. An Excel spreadsheet with VBA interface played each sound in a pre-defined order and collated the responses. For each category, they were first asked to imagine sounds that contained the highest and least 'amounts' of that category. These 'amounts' were given values of 5 and 0 respectively. On hearing each sound, the subject rated the

sound between the fixed markers 0 and 5 – or gave 'No assessment possible' if they felt the category under test did not apply to the assessed sound.

## Presentation of Results and Discussion

Following on from the discussion on reliability, the test method does not distinguish the reasons why some subjects rate a sound at 4 while others rate a sound at 3. The measurement scale is not known to be a scalar in which the difference between rating 0 and 1 is not necessarily the same as between ratings 4 and 5. What is known is that, by virtue of making the decisions about what rating to apply to a sound, the subject is ranking sounds in an ordinal scale. Most of the results presented here are based on this ordinal ranking rather than the absolute judgements.

## Test Reliability

As an initial indication of the reliability of the tests, the assessments from Group A and the assessments from group B have been averaged for each of the 12 attribute categories. These mean ratings are plotted in Figure 1 and show some evidence of agreement between the assessments from subjects in each of the groups.

Similar results can be obtained from Cronbach's Alpha reliability coefficient (Table 1). This coefficient has values in the range 0.0 to 1.0 and provides a measure of inter-rater reliability; in other words the degree of agreement between two or more subjects. An often-quoted Alpha reliability threshold of less than 0.7 is used to indicate unreliable measures. The highlighted scores of less than 0.7 for 'How Thin' and 'How Muted' (group A reliability) and Has Clarity (group B reliability) are indicative that the relevant attributes are not providing the expected consistent response. In the case of 'How Thin' estimates, the graph in Figure 1 shows that the full range of attribute scores is not being displayed.
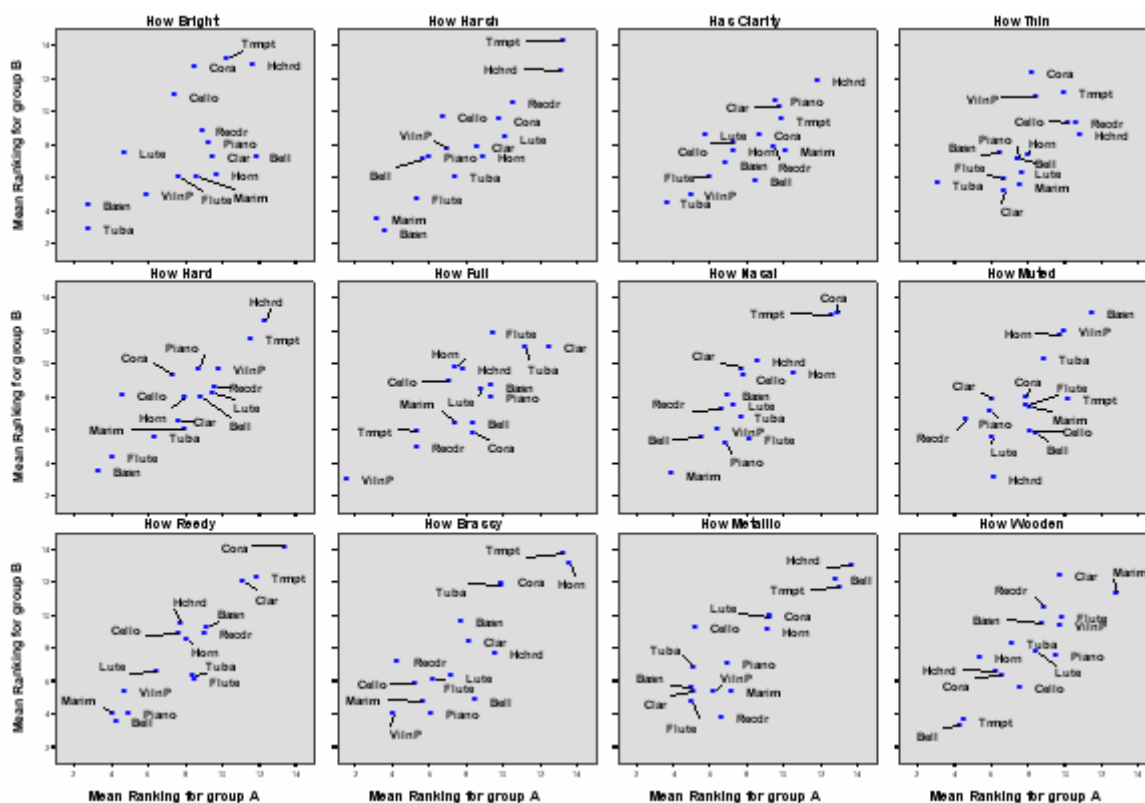


**Figure 1.**  Mean Assessments for Group A vs Group B subjects

| Category | Alpha rating Group A | Alpha rating Group B | Alpha rating All assesors |
|---|---|---|---|
| How Bright | .8758 | .9204 | .9296 |
| How Harsh | .9061 | .9096 | .9519 |
| Has Clarity | .7316 | .6864 | .8413 |
| How Thin | .6239 | .7362 | .8029 |
| How Hard | .8228 | .8058 | .9008 |
| How Full | .8187 | .8167 | .8976 |
| How Nasal | .8035 | .8596 | 9135 |
| How Muted | .6042 | .8597 | .8637 |
| How Reedy | .8770 | .9213 | 9500 |
| How Brassy | .8965 | .9413 | .9544 |
| How Metallic | .9057 | .8968 | .9461 |
| How Wooden | .7767 | .8590 | .9041 |

**Table 1.**        Cronbach's Alpha Reliability Coefficient

One-way ANOVA tests have also been performed between groups A and B to show the level of agreement for individual sounds (Table 2). The F ratio is a measure of the variance between groups compared to the variance within each group. Dependent on the number of subjects (the degrees of freedom), this provides a level of confidence on whether the results are significantly difference other than by chance alone.

These results show that, between groups A and B, there are significant differences in the assessments for the Bell, Cello, Cor Anglais, Horn and Trumpet sounds for 'How Bright', for the Bell, Cor Anglais, Piano and Recorder sounds for 'How Brassy' and the Flute for 'How Brassy'.

| Category | Instrument | F ratio | Significance |
|---|---|---|---|
| How Bright | Basn | .421 | .523 |
| | Bell | 5.756 | .026 |
| | Cello | 4.929 | .038 |
| | Clar | 3.686 | .069 |
| | CorA | 11.140 | .003 |
| | Flute | .823 | .375 |
| | Hchrd | .938 | .344 |
| | Horn | 4.353 | .049 |
| | Lute | 3.381 | .080 |
| | Marim | 2.983 | .099 |
| | Piano | 1.389 | .252 |
| | Recdr | .001 | .971 |
| | Trmpt | 6.672 | .017 |
| | Tuba | .112 | .742 |
| | VilnP | .377 | .546 |

| Category | Instrument | F ratio | Significance |
|---|---|---|---|
| How Brassy | Basn | 1.656 | .212 |
| | Bell | 10.919 | .003 |
| | Cello | .131 | .721 |
| | Clar | .134 | .718 |
| | CorA | 4.917 | .038 |
| | Flute | .049 | .827 |
| | Hchrd | 1.133 | .299 |
| | Horn | .694 | .414 |
| | Lute | .732 | .402 |
| | Marim | 1.059 | .315 |
| | Piano | 7.008 | .015 |
| | Recdr | 8.705 | .008 |
| | Trmpt | .114 | .739 |
| | Tuba | 1.262 | .274 |
| | VilnP | .132 | .720 |

| Category | Instrument | F ratio | Significance |
|---|---|---|---|
| How Full | Basn | .276 | .605 |
| | Bell | .870 | .362 |
| | Cello | .735 | .401 |
| | Clar | 2.884 | .104 |
| | CorA | 4.010 | .058 |
| | Flute | 4.821 | .039 |
| | Hchrd | 1.441 | .243 |
| | Horn | 1.942 | .178 |
| | Lute | .072 | .792 |
| | Marim | .159 | .694 |
| | Piano | 1.100 | .306 |
| | Recdr | .076 | .786 |
| | Trmpt | .012 | .915 |
| | Tuba | .067 | .798 |
| | VilnP | 2.019 | .170 |

**Table 2.**        One-way ANOVA between assessments for groups A and B[1]
(How Bright, How Brassy and How Full only)

## Measurement Validity

If each subject used the same definition of an attribute when making subjective ratings, the

[1] For clarity of presentation, only ANOVA results for 'How Bright', 'How Brassy' and 'How Full' have been included

expectation is that the ratings would cluster around the same regions of timbre space. Cluster Analysis highlights these groupings as shown for 'How Bright' assessments in Figure 2. In this example, subjects RASS16 – RASS04 form one cluster, RASS15 – RASS20 form a second cluster and subject RASS07 is an outlier.
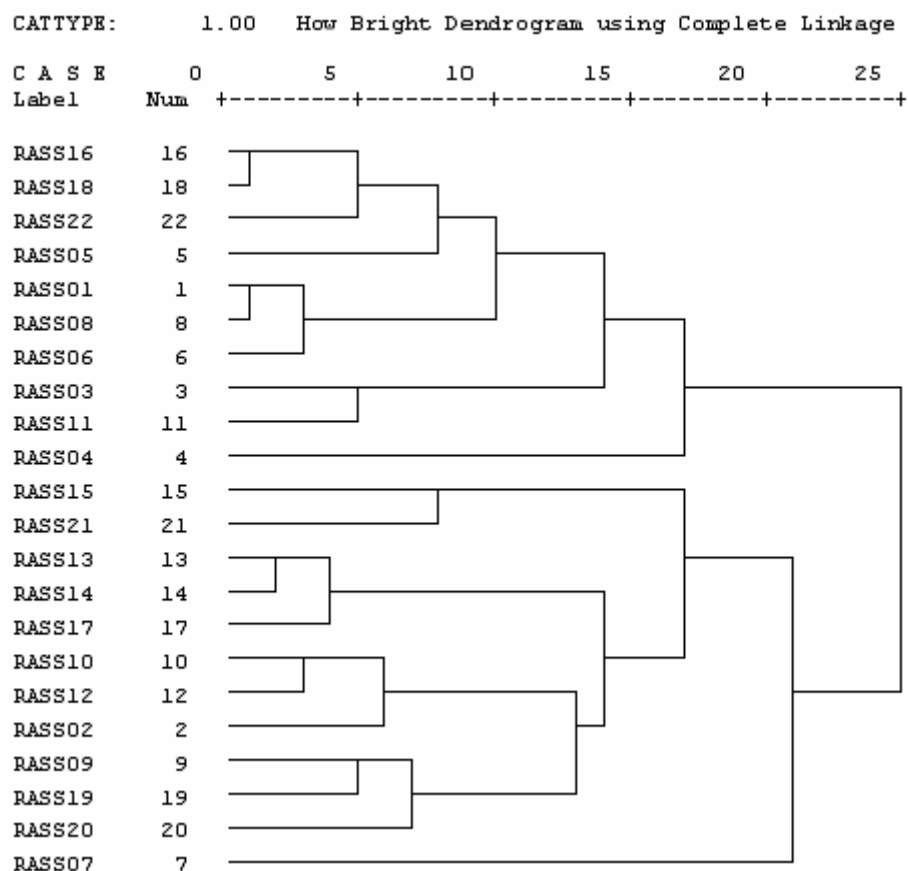


```
CATTYPE:        1.00    How Bright Dendrogram using Complete Linkage

C A S E        0         5        10        15        20        25
Label        Num  +---------+---------+---------+---------+---------+

RASS16        16
RASS18        18
RASS22        22
RASS05         5
RASS01         1
RASS08         8
RASS06         6
RASS03         3
RASS11        11
RASS04         4
RASS15        15
RASS21        21
RASS13        13
RASS14        14
RASS17        17
RASS10        10
RASS12        12
RASS02         2
RASS09         9
RASS19        19
RASS20        20
RASS07         7
```

**Figure 2.**          Cluster Analysis of subject ratings for 'How Bright' assessments

## Test Sensitivity

Categories and sounds that provide a sensitive measure of a subject's attitude show a tendency towards particular rankings with low variance. Histogram plots show graphically the spread of results for each instrument and count the number of occurrences of each assessment ranking (Figure 3). Similar statistics – the variance and kurtosis statistics – give a measure of the spread of observations and have been tabulated in Table 3 for the brightness and brassiness attributes. The highlighted cells, for which the variance is low, appear to be the most consistently ranked (bassoon, harpsichord, piano and tuba for brightness, clarinet, horn, marimba, piano, trumpet and plucked violin for brassiness).

From the histogram of 'How Brassy' rankings (Figure 3), the horn and trumpet are consistently rated as very brassy. It is not clear from the experimental method, and with knowledge of the instruments, whether the subjects are rating the instrument rather than the sound.

Interestingly, if the variance of all instrument rankings are plotted against mean assessment ranking, it indicates that sounds at the extreme ranges of attribute assessments provide a more consistent ranking (Figure 4). Subjects appear more likely to agree on which sounds are most bright and least bright than they are to agree on the ranking of middle order sounds.
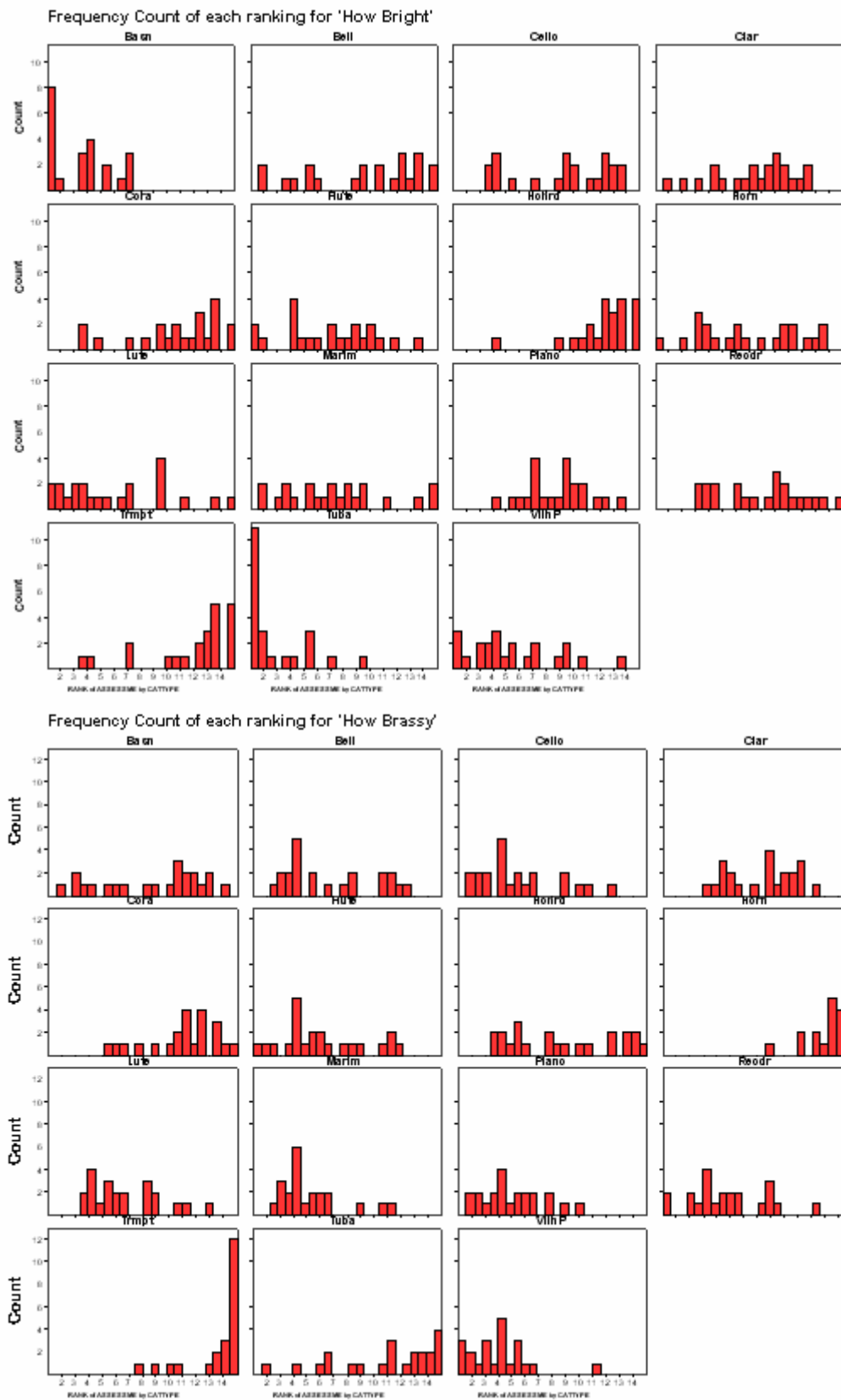
**Figure 3.**  Frequency count of rankings by instrument for 'How Bright and 'How Brassy'

| Category | Instr | Mean Statistic | Variance Statistic | Kurtosis Statistic | | Category | Instr | Mean Statistic | Variance Statistic | Kurtosis Statistic |
|---|---|---|---|---|---|---|---|---|---|---|
| How Bright | Basn | 3.59091 | 4.515 | -1.255 | | How Brassy | Basn | 8.72727 | 14.422 | -1.212 |
| | Bell | 9.59091 | 17.801 | -1.038 | | | Bell | 6.70455 | 11.325 | -1.354 |
| | Cello | 9.25000 | 12.661 | -1.252 | | | Cello | 5.56818 | 8.745 | -.024 |
| | Clar | 8.40909 | 9.063 | -.408 | | | Clar | 8.36364 | 5.957 | -1.225 |
| | CorA | 10.63636 | 11.195 | .152 | | | CorA | 10.97727 | 6.845 | -.152 |
| | Flute | 6.88636 | 11.570 | -.695 | | | Flute | 6.15909 | 9.676 | -.725 |
| | Hchrd | 12.29545 | 5.444 | 5.109 | | | Hchrd | 8.68182 | 15.108 | -1.527 |
| | Horn | 8.00000 | 13.381 | -1.227 | | | Horn | 13.40909 | 1.991 | 3.581 |
| | Lute | 6.11364 | 15.617 | -.572 | | | Lute | 6.79545 | 7.040 | -.113 |
| | Marim | 7.40909 | 14.420 | -.367 | | | Marim | 5.27273 | 5.732 | 1.568 |
| | Piano | 8.75000 | 5.423 | -.454 | | | Piano | 5.09091 | 4.944 | -.277 |
| | Recdr | 8.93182 | 10.412 | -1.170 | | | Recdr | 5.79545 | 8.135 | -.008 |
| | Trmpt | 11.77273 | 11.208 | 1.045 | | | Trmpt | 13.50000 | 4.786 | 2.005 |
| | Tuba | 2.86364 | 5.457 | 1.761 | | | Tuba | 10.88636 | 14.451 | -.116 |
| | VilnP | 5.50000 | 10.667 | .102 | | | VilnP | 4.06818 | 4.888 | 3.401 |

☐ Assessments with low variance for brightness and brassiness

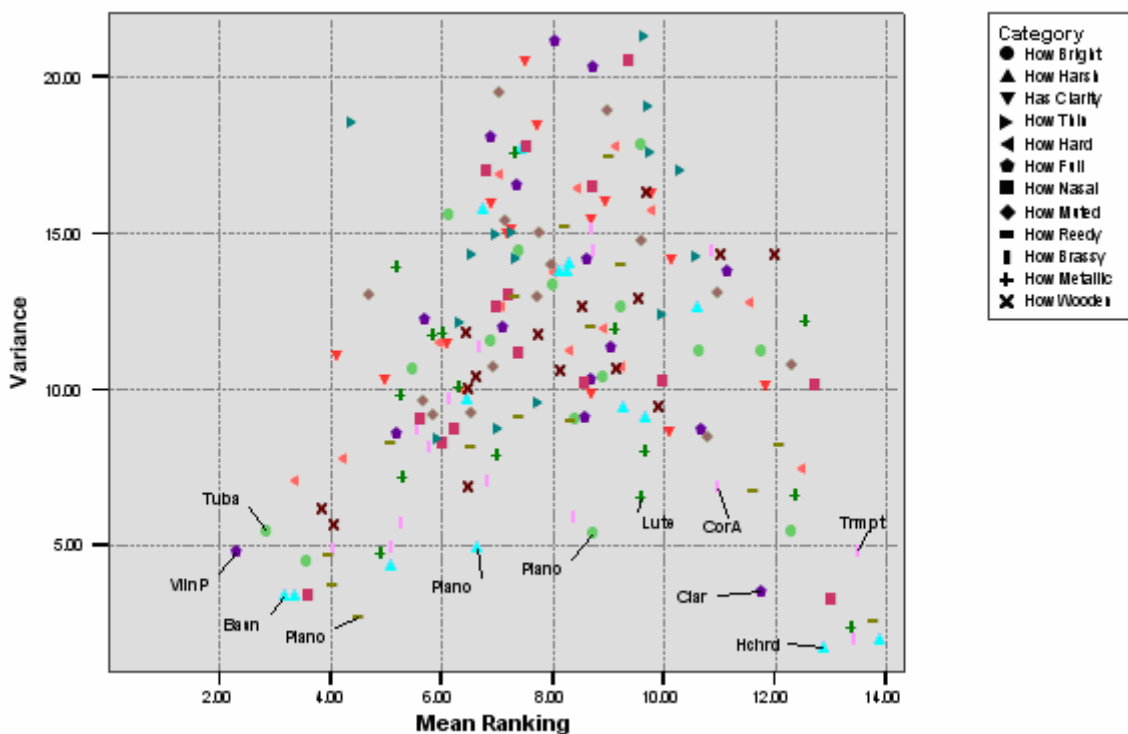**Table 3.**      Variance and Kurtosis statistics for distribution of rankings for Brightness and Brassy



**Figure 4.**      Plot of Variance against Mean Ranking

## Conclusions

The graphs shown in Figure 1 appear to show that there is some evidence of agreement between assessors. However, there is significant variance in the responses from different subjects (Figure 4). The more difficult a category or sound is to rate, the more variance is to be expected between the assessments; Brassiness has best record; Thinness has worst record. The reasons behind the

assessment variances for each category warrant further research to discover if these are due to assessment errors (measurement validity), fuzziness in deciding on a rating (reliability) or some other cause.

The general answer to the question

> '*Can a subject intuitively arrive at consistent attribute ratings on hearing a sound once?*

appears to be '*No*'. Consistency may only be measured with multiple trials. The response to this may be taken two ways. The first response is to repeat the same trial on the same users and look for test-retest reliability. There are difficulties in this approach since the subjects are unlikely to repeat the test using the same strategy since each sound will then be heard again and there will be an element of subject training; responses cannot then be intuitive. The second response is to modify the question to what was actually tested here:

> *Will multiple subjects rate the same sound similarly, despite having only heard the sound once?*

The results show that some sounds are rated by multiple subjects with less variance than other sounds in some categories. This suggests that some sounds have greater significance in the rating of particular categories. What has not been studied here is whether there is any correlation between the sounds assessed with a 'No assessment possible' response from some subjects and a larger combined variance from other subjects.

If there is more than one understanding of a category's interpretation, assessments cluster into different groupings.

Some subjects provide responses that are more consistent than others.

There were differing views on whether to apply 'No Assessment Possible' or to best guess an appropriate reply for difficult sounds (e.g. how do you rate a harpsichord for brassiness). With more sounds and more subjects, one would expect a trend to occur in results for different variances.

Stimulus presentation order appears to influence responses between groups A and B for some sounds.

The discussion in this paper has shown that there is no conclusive evidence that subjects agree on how to effectively communicate timbral issues. The paper has highlighted some of the issues behind why people do not completely agree, but that some trends do exist when comparing results.

## What Next?

After the execution of this experiment, several subjects made comments that they had made a few assessment errors. The opportunity to retest would allow subjects to validate their initial assessments. The objective is to verify whether assessment reliability improves on a controlled retest. The following criteria is to be used:

- In this experiment, subjects move the sounds into different bins. Each 'bin' contains only sounds of the same ranking. Therefore, stimulus presentation order is not important. Are there still significant differences in some sounds between groups A and B?
- Since subjects rank only the sounds within the test, no external 'imaginary' reference points are used.
- Assessment granularity is self-generating since subjects can create as many bins as there are 'Just Noticeable Differences' between levels.

The evidence for both Musch & Klauer's postulates (above) stem in a large part from tests using affective priming. In the studies looked at by Musch & Klauer, the evaluation of polarized target words proceeds faster and more accurately when preceded by an evaluatively consistent prime word. The majority of the studies looked at were based around the cognition of language. Is there a similar pre-conscious process involved in timbre when a short, neutral 'priming' sound is heard immediately prior to the target sound.

It is anticipated that the primary link between the acoustical structures of a musical sound and the perceptual domain will follow on from the confirmation by Creasey and others that '*Timbre*

*space…can be represented by grouped features derived from time and frequency representations'* (Creasey, 1998, chapter 6.5). An example of this representation is the tristimulus method of Pollard and Jansson (Pollard and Jansson, 1982).

These studies have looked primarily at using physical measurements to investigate which parts of a sound are significant in auditory perception. By careful selection of aspects of the time and frequency representation, various methods can be used to either predict verbal attribute assessments, or to generate new sounds matching certain attribute criteria.

In summary, sets of qualitative assessments are taken of a series of attributes of a sound. Based on using verbal attribute magnitude estimation (VAME), any correlation between the acoustical analysis and these VAME assessments could provide an insight into how we perceive and understand the component attributes of timbre.

# References

Begault, D.R. (2003). Experimental Design – an Introduction. Tutorial on Listening Tests 115th AES Convention, New York NY, October 2003.

Creasey, D.P. (1998). An exploration of sound timbre using perceptual and time-varying frequency spectrum techniques. PhD Thesis, University of York.

Grey, J.M., Gordon, J.W. (1978). Perceptual Effects of Spectral Modifications on Musical Timbres. Journal of the Acoustical Society of America, 63 (5), p. 1493-1500.

Kendall, R.A., Carterette, E.C. (1993). Verbal attributes of simultaneous wind instrument timbres: 1. Von Bismarck Adjectives. *Music Perception* 10(4) p. 445-468.

Musch J., Klauer K.C., eds. (2003). The Psychology of Evaluation. Lawrence Erlbaum Associates.

Osgood G. et al. (1957). *The Measurement of Meaning*. University of Illinois Press, Urbana 1957.

Pitt, M.A. (1995). Evidence for a Central Representation of Instrument Timbre. *Perception & Psychophysics*, 57 (1), p. 43-55.

Pollard, H.F., Jansson, E.V. (1982). A Tristimulus Method for the Specification of Musical Timbre. *Acustica,* Vol 51 (1982), p. 162 – 171.

Wake, S. & Asahi, T. (1998). Sound Retrieval with Intuitive Verbal Expressions. *ICAD'98.*