Audio Engineering Society

Convention Paper 8837

Presented at the 134th Convention
2013 May 4–7 Rome, Italy

A pairwise and multiple stimuli approach to perceptual evaluation of microphone types

Brecht De Man[1], Joshua D. Reiss[1]

[1] *Centre for Digital Music, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom*

Correspondence should be addressed to Brecht De Man (brecht.deman@eecs.qmul.ac.uk)

ABSTRACT

An essential but complicated task in the audio production process is the selection of microphones that are suitable for a particular source. A microphone is often chosen based on price or common practices, rather than whether the microphone actually works best in that particular situation. In this paper we perceptually assess six microphone types for recording a female singer. Listening tests using a pairwise and multiple stimuli approach are conducted to identify the order of preference of these microphone types. The results of this comparison are discussed, and the performance of each approach is assessed.

1. INTRODUCTION

The selection of microphones is a crucial phase of the audio production process. Among other things, it determines the frequency response with which the sound source is filtered before being amplified and (in a digital console or digital audio workstation) converted to the digital domain. Whereas equalising can in theory shape an almost arbitrary amplitude response, recording and mixing engineers usually start the signal chain with a high quality signal that requires as little processing as possible. This not only simplifies the music production process and reduces the necessary processing overhead, but also avoids the artefacts that (heavy) processing may induce. Another part of the characteristic sound of a microphone is due to non-linear characteristics, which can either be desired or undesired and which can not be compensated easily by linear processing such as equalising. Choosing a microphone that suits the source and envisaged sound as closely as possible is therefore vital. Any bias due to knowledge of the microphones' cost and common practices can be detrimental to the accuracy of the engineer's judgement. In this paper we investigate whether different subjects show a consistent microphone preference within a range of microphone types and prices,

solely based on blind subjective evaluation. As an example we consider the situation where six different microphones are available for the recording of a female singer[1]. An ulterior goal is to compare the performance of pairwise and multiple stimuli evaluation for this type of task.

2. SETUP AND RECORDING CONSIDERATIONS

A selection of six commonly used microphone types were tested, including condenser microphones from entry level (Audio Technica 2020) to professional level (AKG C414 B-XL II), as well as dynamic instrument and vocal microphones (Shure SM57, Shure Beta 58A and Electro-Voice RE-20) and a ribbon microphone (Coles 4038). They were arranged closely together, at equal distance (± 30 cm) from the singer's mouth, thus minimising variations in distance and phrasing by allowing for simultaneous recording. Good recording practice was followed to the greatest possible extent for each of them (alignment of the microphone axis with the singer). It should be noted that every microphone has its own 'sweet spot', meaning that it is advisable to use some microphones closer to the singer's mouth, also depending on the source and the desired sound. However, this distance has been chosen as an appropriate average for this experiment. To reduce the effect of the variation in directionality and the corresponding proximity effect as well as the ratio of direct and reflected sound, all microphones were set to cardioid where possible. Exceptions are the Shure Beta 58A, which is hypercardioid by default, and the Coles 4038, which like most ribbon microphones has a figure-eight characteristic. Note that all of these microphones would exhibit the proximity effect as they are all directional, albeit not all with the exact same directionality pattern. The microphones under test are listed in Table 1, together with the numbering that is maintained throughout this text.

The recording took place in the Listening Room at the Centre for Digital Music at Queen Mary University of London. The microphone preamp used was a Focusrite ISA828, and it was set so that the input levels before digitisation were roughly equal, when

[1]Eight microphones were originally tested, but results on two microphones have been excluded due to errors in the acquisition of signals from those two microphones.

M1	Audio Technica AT2020	condenser
M2	AKG C414 B-XL II	condenser
M3	Coles 4038	ribbon
M4	Shure SM57	dynamic
M5	Shure Beta 58A	dynamic
M6	Electro-Voice RE-20	dynamic

Table 1: Microphones under test.

Fig. 1: The singer in front of the microphone array.

excited by pink noise through a loudspeaker. The SSL Alpha-Link MADI AX provided the analog-to-digital conversion (ADC). The files were recorded in 24 bit, 44.1 kHz WAV format. After this, the audio files were not processed in any way.

We chose the human voice as a source, because people are known to be able to discriminate very subtle differences in the sound of the human voice [1]. Playback of the test sound through a loudspeaker was considered undesirable here because the transfer function of the original microphone and of the loudspeaker itself would render the situation different from reality. Rather, we chose to conduct an experiment that would resemble a real-life microphone comparison in a recording studio to the greatest possible extent.

The singer was asked to sing fragments of a loud, high-pitched rock song and a softer, low-pitched jazz song (in this case *Black Velvet* and *No More Blues (Chega de Saudade)*, respectively). A clean, four-second fragment was chosen as test material, of which the lyrics are "Black velvet and that little boy's smile", and "There'll be no more blues". Notice the absence of aspirated plosives ('popping' sounds). Because there were no significant popping artefacts in these clips, no pop filter was used, although the same material recorded with pop filter was also available. To ensure proper flow and phrasing, as well as a sufficient amount of data, whole choruses were sung, whereas only a short clip was used for the tests.

Rather than a frequency response obtained by recording a static, spectrally flat source (relative to a reference microphone), Figures 2 and 3 show the spectra of the two samples for each microphone. Notice the relative difference between the microphones depending on which sample excites them.

To minimise perceived loudness differences, which could cause a bias towards louder (or softer) samples, the loudness of every sample was equalised using the subjective programme loudness calculation in [2].

3. LISTENING TEST

The listening test for this task presents a few challenges. Because of the strong similarity of the samples, a multi-stimulus test where all samples are compared at the same time appeared to be quite hard during the pilot tests. Pairwise comparison, on the other hand, allows the subject to examine the differences between the samples more accurately. However, since 21 comparisons are needed for as little as six different samples, this approach can strain the subject's attention considerably. Thus, both approaches are applied and compared here.

The only question asked in this test is deliberately chosen to be a very subjective and general one: to rate the perceived quality of the different samples compared to another one (pairwise test) or compared to every other one at the same time (multi-stimulus test). As such, the subjects are not asked explicitly which sample sounds the most accurate (subjects sometimes prefer a 'distorted' version of a sound when no reference is available [1]), or desirable in a recording context (which would also require the subjects to have audio engineering experience). The subject's main impression of the programme material relates solely to timbre, freedom from noise and distortions, and - to some extent - transparency, as there is only one source (sound balance is no factor) and the recordings are mono (no variations in stereo impression) [3]. Because the room used for recording was fairly dry, and the microphones were all close to the source, the variations in spatial impression are small too.

The listening test is conducted using two sets of short audio clips, recorded with the aforementioned configuration, in a quiet room using a pair of high quality, circum-aural headphones. The transfer function of the headphones being used (Beyerdynamic DT770 Pro) is displayed in Figure 4.

One group of subjects evaluates the first set of samples in a pairwise fashion, establishing the order of their preference by means of side-by-side comparison of every possible microphone pair, and testing their reliability by including pairs of the same microphone. The subjects are aware of the fact that some paired samples may be exactly the same. However, they do not know that the reason the samples sound different is because they are recorded by different microphones. The second set of samples is then evaluated by presenting all samples at once, allowing the test subject to place each sample on a one-dimensional near-continuous axis to reflect their overall, subjective impression. A second group evaluates the first set of samples using the multiple sti-

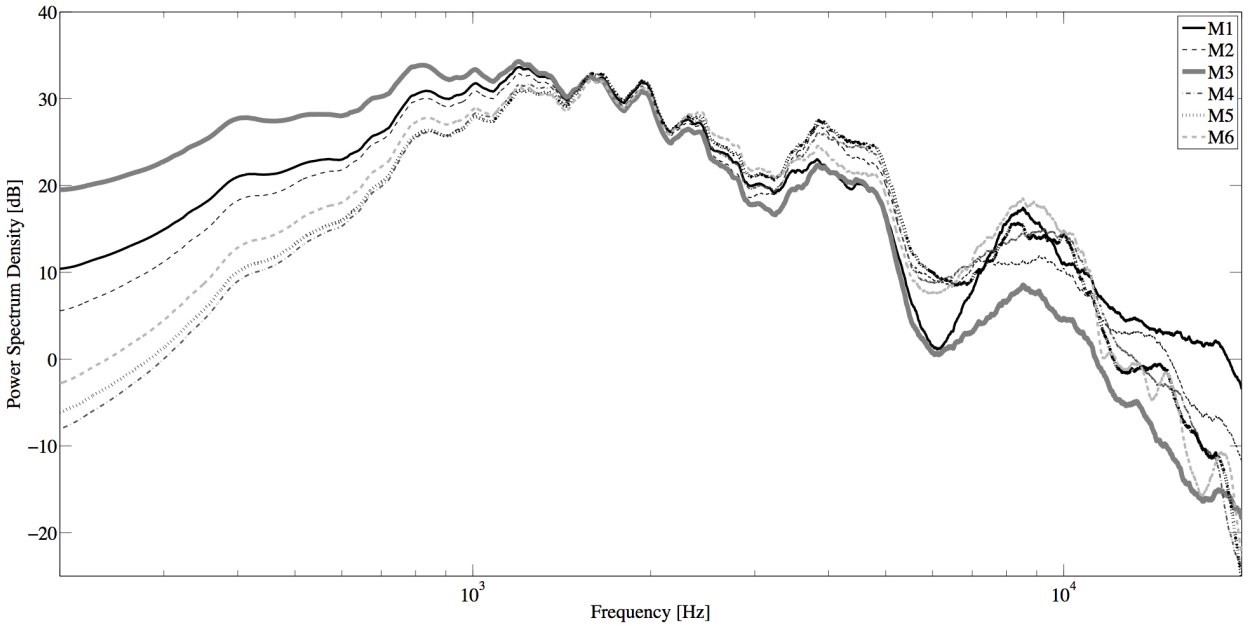

Fig. 2: Power spectra of the "Black Velvet" samples.

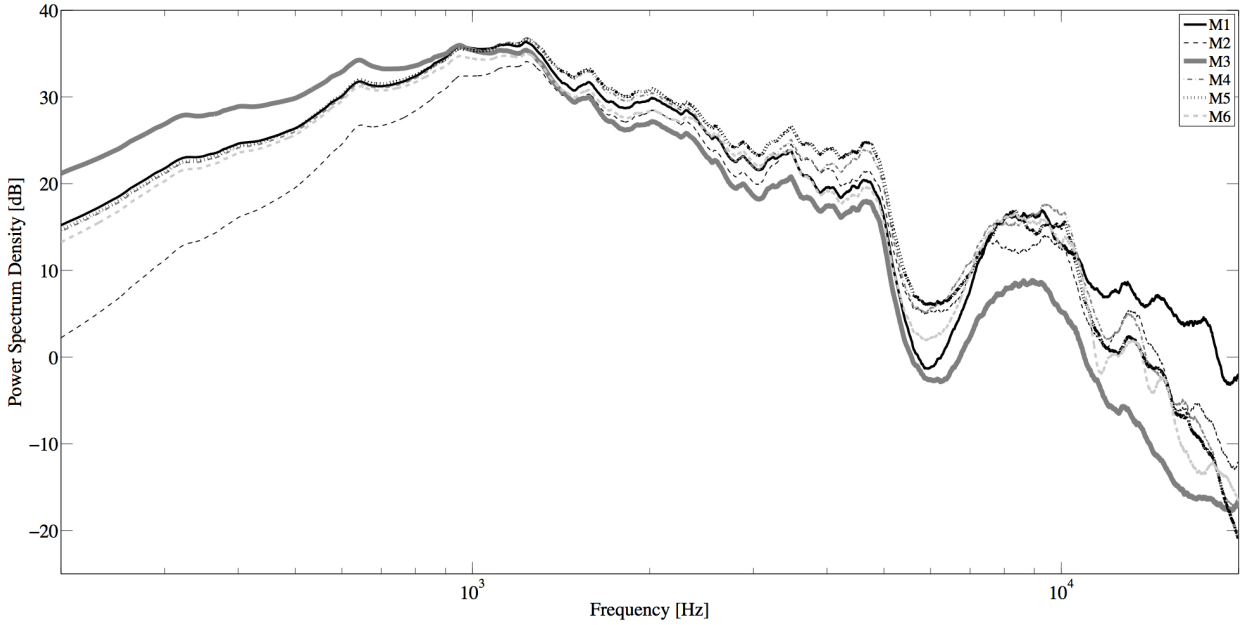

Fig. 3: Power spectra of the "No More Blues" samples.

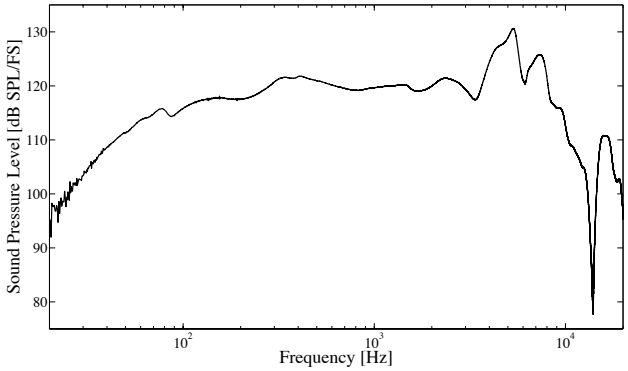

Fig. 4: Transfer function of the set of headphones used for the listening test, as measured using a KEMAR artificial head and sine sweep excitation. It is an average of three left and right channel recordings, and shows the SPL at 0 dBFS in function of frequency.

muli approach, and the second set of samples using the pairwise approach. To remove the extra bias due to always assessing one particular set of samples first (or always starting with the same type of listening test), the order of these modules is also alternated throughout the experiments. This means that in each group, half of the participants assess the first set of samples first and then the second set of samples (using said approaches), while the other half does the opposite (see Table 2 for a complete overview of the subject groups).

Table 2: Subject groups

Group	Tests
1.1	Pairwise "Black Velvet" Multiple Stimuli "No More Blues"
1.2	Multiple Stimuli "No More Blues" Pairwise "Black Velvet"
2.1	Multiple Stimuli "Black Velvet" Pairwise "No More Blues"
2.2	Pairwise "No More Blues" Multiple Stimuli "Black Velvet"

Before the actual test, the subjects received written and verbal instructions [4], and were asked to play back and compare a set of trial samples (presented in randomised order), in order to familiarise themselves with the material (as well as with the multiple stimuli listening test interface, which was used for this familiarisation stage). The trial samples were different from the samples used in the test, but were recorded in the exact same way so as to demonstrate the differences in quality the subjects could expect throughout the experiment.

After each test, the subjects are asked to complete a short questionnaire about their experience as a musician (years of experience playing one or more instruments), experience as an audio engineer (recording, mixing or other technical audio tasks), experience with listening tests and whether or not they have hearing impairments, or a cold or ear infection that may have affected their hearing. Subjects who had experience as an audio engineer were asked if they hypothetically could recognise certain microphones, to determine if this could potentially bias their choices.

3.1. Pairwise comparison

For pairwise comparison, 21 questions are necessary per test. This is because every sample is compared to itself once (1-1 is a valid combination), and no comparison is repeated (1-2 is never presented when 2-1 is, and vice versa). This leads to a number of combinations where order is important and repetition is allowed, calculated as

$$\frac{(n+r-1)!}{r!\,(n-1)!} = \frac{(6+2-1)!}{2!\,(6-1)!} = 21 \qquad (1)$$

The order of the presented pairs is randomised to remove as much bias as possible.

The only question in this test is whether the subject prefers A or B, or if there is no perceivable difference. The latter can be the case when the samples are, in fact, equal, or when they are not but the subject doesn't hear a difference. This is one of the ways to monitor the subject's reliability. They are encouraged not to claim the samples are the same if they do hear a difference, but do not have a clear preference for one or the other.

The pairwise interface simply shows two buttons to play the respective audio clips, that can be clicked

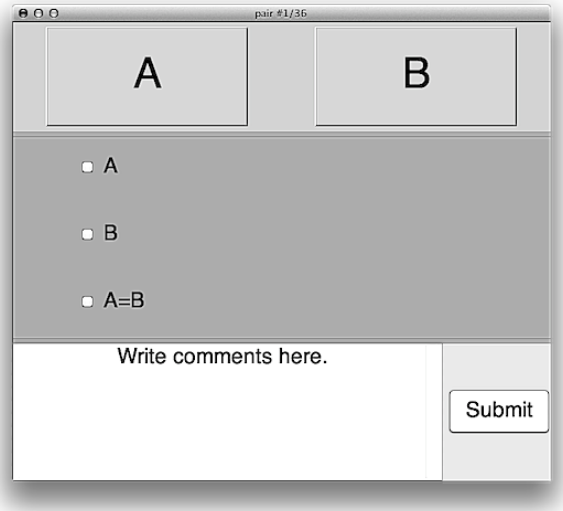

Fig. 5: A user interface similar to the one used for the pairwise evaluation.

indefinitely and without delay, tick boxes with the options 'A', 'B' and 'A=B', an optional comments field, and a 'Submit' button (see Figure 5). The subjects are instructed to choose 'A' in case sample A sounds best to them, 'B' if sample B sounds best, and 'A=B' in case they do not hear a difference.

3.2. Multiple stimulus comparison

A common listening test type for the subjective evaluation of audio samples is MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) [4, 5, 6, 7, 8]. However, unlike with audio codec tests, there is no 'reference' among the samples, so we choose not to use this term and do not require one score to be maximum, as is the case with MUSHRA. An example of this can be found in [9]. Therefore, the methodology for this experiment is different from a MUSHRA test. Furthermore, to encourage thorough comparison of the samples, it has been the author's choice to let the subject position the audio samples on one one-dimensional scale, rather than to rate them with 6 independent sliders. For our purpose, the goal of this test is to obtain a ranking of the different samples, rather than an opinion score. It has been shown that ordinal scales (rankings) can be preferable to interval scales (numerical ratings) for listening tests [10].

The main reason for using a MUSHRA-style approach, where subjects can rate different samples on a (near-)continuous scale, is that this allows for rating of very small differences. This is easily understood when one imagines a mean opinion score (MOS) test, where subjects could rate 4 out of 6 microphones as 'Good', providing very little information, which would require many participants to obtain statistically significant results. The single, one-dimensional, near-continuous scale on which the 6 markers are to be positioned simplifies the task of assessing the samples and provides us with more information. For example, as opposed to a plain ranking interface, we can now learn which samples a subject perceives as almost equal, in case we would like not to incorporate this 'ranking' (e.g. 1 is only slightly better than 2) in the analysis.

For the interface, a single bar with a marker per microphone is displayed, which can be clicked to play back the corresponding sample - again as often as desired - and dragged around to indicate the preference on an unmarked scale. Initially, the markers are scattered randomly across the bar. To make this task easier for the subject, the markers each display a number from 1 through 6 (see Figure 6). This number is also randomly assigned and as such does not correspond with the numbering of the microphones, to remove any bias this could induce (as subjects might tend to listen to the different clips in order, i.e. from 1 to 6).

4. STATISTICAL ANALYSIS AND RESULTS

The sample size for both the pairwise and the multiple stimuli test is 36, as each of the 36 subjects participated in both tests. Each of the 4 groups (see above) contains 9 subjects.

To make direct comparison between both methods possible, a list of all possible A/B pairs (with A different from B) is constructed, along with the times A was chosen over B, and vice versa. For the pairwise comparison, this is a very straightforward task, as the answers of all 36 participants can simply be summed. If a pair of two different microphones is labeled as equal (i.e. the subject was not able to distinguish between the samples), no 'votes' are added.

In case of the multiple stimuli test, the order of the markers is easily converted to such a list too. This

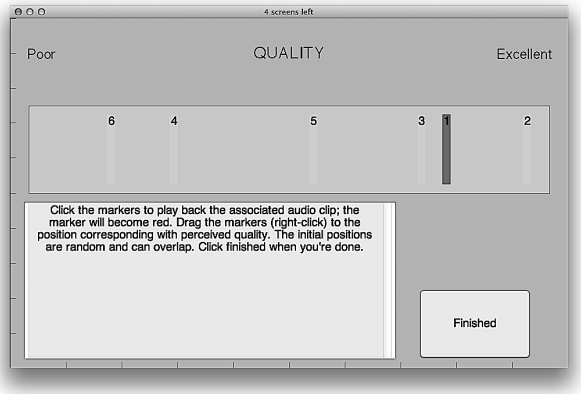

Fig. 6: A user interface similar to the one used for the multiple stimuli evaluation. The marker corresponding with the clip that was played last is always highlighted.

means that if microphone M1 is on top, all combinations containing M1 get a 'vote' in favour of M1. If M3 is rated second, all combinations containing M3 get a 'vote' in favour of sample M3, except for combination M1-M3, in which M1 is the one that was chosen. One exception is made here: when two or more markers are put at the same position on the quality scale, they are assumed to be rated equal. This means that if M1 and M2 are both positioned exactly halfway the range by the subject, no scores are added for the M1-M2 pair.

To decide if the number of subjects who preferred A over B is significantly different from the number of subjects that preferred B over A, the normal approximation is used as follows [11]. First, we calculate the *z-score*:

$$z_{sample} = \frac{(P_s - 0.5) - P_u}{\sqrt{\frac{P_u(100-P_u)}{n}}} \quad (2)$$

where P_s is the *sample percentage*, P_u is the *assumed population percentage*, and n is the sample size. To obtain the sample percentage needed to consider A and B significantly different in average perceived quality, meaning that the alternative hypothesis, $H_A : P_s > 50\%$, is true as opposed to the null hypothesis, $H_0 : P_s = 50\%$, we rearrange the equation as follows:

$$P_s = z_{sample}\sqrt{\frac{P_u(100 - P_u)}{n}} + P_u + 0.5 \quad (3)$$

The value for z for $p = 0.05$ one-tailed (using the standard significance level of 5%) is 1.64, yielding

$$P_s = 1.64\sqrt{\frac{50(100 - 50)}{36}} + 50 + 0.5 = 64.17\% \quad (4)$$

for a sample size $n = 36$. However, this value will only be accurate if the number of votes for A plus the number of votes for B equals 36, which will rarely be the case as pairs can also be inaccurately called equal, where the subject does not cast a vote for either A or B in the case of pairwise comparison. The same goes for multiple stimuli comparison, when subjects attribute the same quality rating to two or more samples.

The number of subjects justifies the normal approximation used in this analysis. Furthermore, even if a significant fraction of the subjects is not considered, for example when analysing the results of certain subgroups or omitting subjects on grounds of errors they made, the effect of violating the assumption that the data conforms a Gaussian distribution is quite small [10].

Ideally, there is a sufficient number of pairs for which subjects have a significant preference for one of its elements, so we could list a few possible rankings with 95% certainty (e.g. the order is either 135462 or 134562, when only M4 and M5 are not significantly preferred over each other).

In order to measure subject reliability, the following data is available for each subject: (1) the number of times a pair of different samples are labeled as equal, i.e. 'A=B' and (2) the number of times a pair of equal samples are labeled as being different, i.e. 'A' or 'B'. Other performance measures can be calculated, such as the number of violations of transitivity in the pairwise case, for example when a subject says he prefers microphone M1 over M2, M2 over M3 and M3 over M1. In the multiple stimuli case, transitivity is automatically fulfilled. Another possible

measure is the distance between the subject's preferences in the pairwise experiment compared to the preferences in the multiple stimuli - however, this is not necessarily something to avoid as it's possible that a different sample leads to a real different order of preferences.

Furthermore, the incorporation of the short questionnaire allows for example to investigate if considering only subjects with substantial experience in music and/or audio engineering leads to different results.

Table 3 shows the number of votes for microphone A over microphone B for the pairwise test. Table 4 shows the same for the multiple stimuli tests.

A	**B**	**#A**	**#B**	**%A**	**%B**	n	P_s
1	2	11	12	47,83	52,17	23	67,60
1	3	32	4	88,89	11,11	36	64,17
1	4	12	10	54,55	45,45	22	67,98
1	5	16	16	50,00	50,00	32	65,00
1	6	14	13	51,85	48,15	27	66,28
2	3	32	4	88,89	11,11	36	64,17
2	4	12	10	54,55	45,45	22	67,98
2	5	14	15	48,28	51,72	29	65,73
2	6	13	8	61,90	38,10	21	68,39
3	4	4	32	11,11	88,89	36	64,17
3	5	8	27	22,86	77,14	35	64,36
3	6	4	31	11,43	88,57	35	64,36
4	5	12	7	63,16	36,84	19	69,31
4	6	8	13	38,10	61,90	21	68,39
5	6	13	14	48,15	51,85	27	66,28

Table 3: Distribution of subjective votes for pairwise test.

P_s shows the percentage of the votes that either microphone A or microphone B should receive to be able to judge that there is a significant preference for one over the other (in function of n). The highlighted rows show the pairs where this is true. It is immediately clear that there is a significant preference against microphone M3 in both tests, which apparently is perceived as lacking high frequency content compared to the other microphones, as can be seen in Figures 2 and 3.

Additionally, in the multiple stimuli case, microphone M4 is preferred over microphone M6, albeit with considerably less significance than for the pairs containing microphone M3.

A	**B**	**#A**	**#B**	**%A**	**%B**	n	P_s
1	2	14	20	41,18	58,82	34	64,56
1	3	35	1	97,22	2,78	36	64,17
1	4	16	19	45,71	54,29	35	64,36
1	5	16	19	45,71	54,29	35	64,36
1	6	12	21	36,36	63,64	33	64,77
2	3	35	1	97,22	2,78	36	64,17
2	4	16	19	45,71	54,29	35	64,36
2	5	18	17	51,43	48,57	35	64,36
2	6	19	16	54,29	45,71	35	64,36
3	4	5	31	13,89	86,11	36	64,17
3	5	4	32	11,11	88,89	36	64,17
3	6	3	33	8,33	91,67	36	64,17
4	5	20	14	58,82	41,18	34	64,56
4	6	23	11	67,65	32,35	34	64,56
5	6	18	15	54,55	45,45	33	64,77

Table 4: Distribution of subjective votes for multiple stimuli test.

For each pair, the values for n (total number of answers) indicate how many subjects were able to distinguish between these two microphones. Since every pair was examined by each of the 36 subjects, $36 - n$ subjects thought A and B were equal.

When summing the results of the pairwise and multiple stimuli tests, no new information is obtained (i.e. only microphone M3 has a significantly lower percentage of votes than any of the other microphones). Looking at the sample sets separately (*Black Velvet* and *No More Blues*), the results change slightly in the multiple stimuli case: for *No More Blues* the slight preference for M4 over M6 found earlier is no longer significant, whereas for the *Black Velvet* sample, there is also a significant preference for M5 over M1 (12 vs. 5 votes) and M5 over M6 (11 vs. 4). Determining the relation between the samples' characteristics and its influence on the results lies beyond the scope of the paper.

The order of the type of tests, which was also alternated in this experiment, did not have any significant effect on the pairwise ratings (no pair other than the ones from Table 3 showed a significant difference in subjective votes). However, in the multiple stimuli case there were a few differences in votes

that became significant. When the pairwise test was done first, there was an additional preference of M5 over M1 (13 vs. 5), M6 over M1 (13 vs. 4) and M4 over M6 (12 vs. 5; also significant in the general multiple stimuli table, see Table 4). When the multiple stimuli test was done first, those three differences were no longer significant, but M2 was preferred over M1 (11 vs. 6).

Considering only subjects with more than ten years of experience playing an instrument ($n = 18$), the result for both the pairwise and multiple stimuli test is only a preference for all microphones over M3. When considering only subjects who claim to have experience in audio ($n = 20$), the result is the same plus a preference of M4 over M5 in the pairwise case (9 vs. 3).

A	**B**	**#A**	**#B**	**%A**	**%B**	n	P_s
1	2	6	6	50,00	50,00	12	74,17
1	3	19	2	90,48	9,52	21	68,39
1	4	9	4	69,23	30,77	13	73,24
1	5	11	10	52,38	47,62	21	68,39
1	6	9	7	56,25	43,75	16	71,00
2	3	19	2	90,48	9,52	21	68,39
2	4	8	5	61,54	38,46	13	73,24
2	5	11	6	64,71	35,29	17	70,39
2	6	6	5	54,55	45,45	11	75,22
3	4	4	17	19,05	80,95	21	68,39
3	5	6	15	28,57	71,43	21	68,39
3	6	3	18	14,29	85,71	21	68,39
4	5	10	4	71,43	28,57	14	72,42
4	6	3	8	27,27	72,73	11	75,22
5	6	5	10	33,33	66,67	15	71,67

Table 5: Distribution of subjective votes for pairwise test of 21 best performing subjects.

Figure 7 shows the number of errors made by each subject: the fraction of pairs consisting of two different samples the subject labeled equal, and the fraction of equal pairs (as a reliability check) the subject thought were different. If we leave out the data of the subjects who either missed 50% or more equal pairs (the subject thought they sounded different) and/or wrongly labeled at least 50% of the different pairs as being equal (the subject didn't hear a difference), and analyse the data in the same way as before with the remaining 21 subjects, we obtain the

A	**B**	**#A**	**#B**	**%A**	**%B**	n	P_s
1	2	6	14	30,00	70,00	20	68,84
1	3	20	1	95,24	4,76	21	68,39
1	4	9	12	42,86	57,14	21	68,39
1	5	9	12	42,86	57,14	21	68,39
1	6	8	11	42,11	57,89	19	69,31
2	3	20	1	95,24	4,76	21	68,39
2	4	10	11	47,62	52,38	21	68,39
2	5	11	10	52,38	47,62	21	68,39
2	6	12	9	57,14	42,86	21	68,39
3	4	4	17	19,05	80,95	21	68,39
3	5	2	19	9,52	90,48	21	68,39
3	6	2	19	9,52	90,48	21	68,39
4	5	10	10	50,00	50,00	20	68,84
4	6	16	5	76,19	23,81	21	68,39
5	6	13	8	61,90	38,10	21	68,39

Table 6: Distribution of subjective votes for multiple stimuli test of 21 best performing subjects.

preference distribution shown in Tables 5 and 6. In the pairwise case, nothing changes (the same pairs have a significant difference in votes). In the multiple stimuli case, however, there is an additional significant preference for M2 over M1 (14 vs. 6), compared to the original multiple stimuli results.

In general, the multiple stimuli test seems to yield more significant preferences than the pairwise test.

Summing the total number of votes for each microphone to obtain an average score along with a confidence interval, we are faced with the following issue: because there is the option of deciding that two samples are the same, even when they are not, some subjects may assign 15 votes (one for each pair of different samples) and some may for example assign only 8 votes in total, because they didn't hear the difference between two different samples in the 7 other cases. For this reason, we introduce the following scoring system:

- If subject X prefers microphone A over microphone B in pair {A,B}, microphone A receives 2 points and microphone B receives 0 points.
- If subject X prefers microphone B instead, A receives 0 points and B receives 2 points.

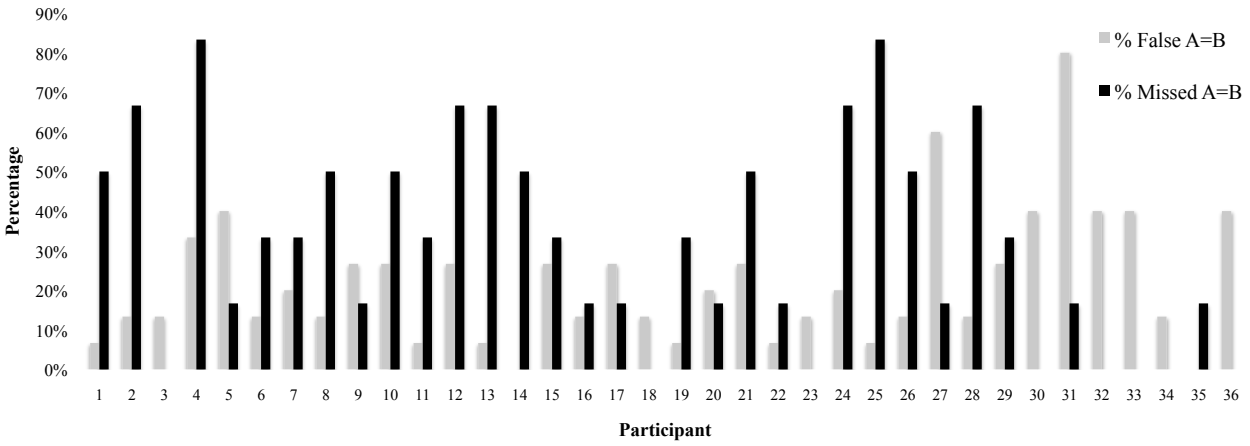

Fig. 7: Fraction of pairs labeled equal when they were not, and pairs labeled different when they were equal, per subject.

- If subject X says A and B are equal, both A and B get 1 point.

That way, every subject distributes 30 points over the 15 relevant pairs, and every microphone gets a score between 0 and 10 points from every subject. For the multiple stimuli test, the same scoring system is used: as before, each multiple stimuli rating is converted to a ranking, which at its turn is converted into a table of A/B pairs, such as the one in Table 4. As some subjects rated some microphones equal (i.e. the corresponding markers were placed within 1% of the full range from each other), this system will allow those microphones to have equal ratings too.

These scores can now be normalised (divided by 10) and analysed, to obtain the means displayed in Figure 8. Again, the only significant result from this is the lower preference for microphone M3. The 0.80 confidence intervals show smaller variation for pairwise comparison, suggesting a higher performance for this method. However, one should be wary of the possibility of this being an artefact of the testing process and statistical comparison, and not due to whether test subjects actually perform better in the case of pairwise testing. This can be understood by considering the following scenario. Suppose all microphones are equal: one could assume that subjects are less likely to award ties for multiple sti-

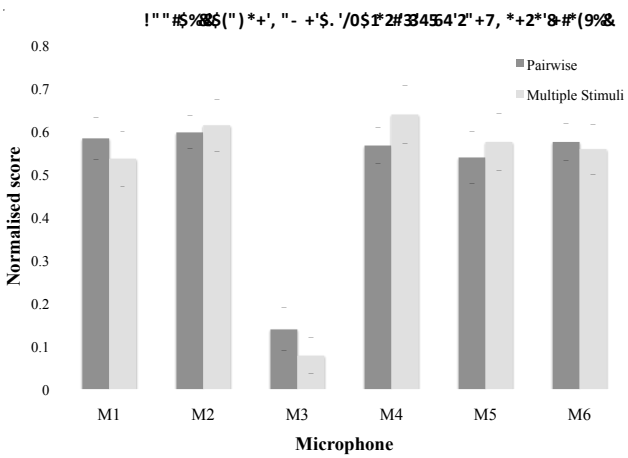

Fig. 8: Preference scores with 80% confidence intervals.

muli tests (an effect that is demonstrated in the next paragraph). In the case of a multiple stimuli test, a microphone would have equal probability of rating 0, 2, 4, ..., 10 (see above). However, in the case of pairwise comparison, scores would tend to award a tie for all microphones - that is, they would each have a far higher chance of rating around 5 than one of the extremes, i.e. 0 or 10. Thus, there would be only little variance for the pairwise case, and a lot of variance in the multiple stimuli case if subjects voted 'randomly', or if they had severe difficulties distinguishing between microphones.

Figure 9 demonstrates the use of the rating scale in the multiple stimuli tests. The average span (the portion of the rating scale used, i.e. maximum rating minus minimum rating per subject) is 62.49% of the scale, with one subject using as little as 7.03%, and another one using the full range (0%-100%). The standard deviation of the span is 22.42%. It is evident from this plot that subjects are far less likely to award the exact same score to microphones than in the pairwise case.

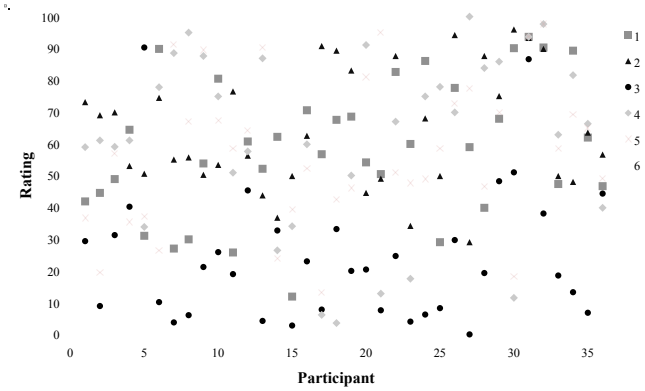

Fig. 9: Rating of microphones on the one-dimensional axis in the multiple stimuli interface.

5. CONCLUSIONS

Out of the 6 microphones that were featured in this test, only one (M3) received a rating that was significantly different (lower) from the other microphones. Depending on the test used and the subgroup considered, some other preferences between pairs of microphones were revealed. Overall, the multiple stimuli test led to more of such preferences, suggesting a better performance. On the other hand, a tighter confidence interval for the pairwise evaluation when summing all scores for each microphone (see Figure 8) suggests a better performance for this method. It is not certain which method most accurately shows the subjects' preferences.

It should be noted that failure to reject the null hypothesis (microphone A is preferred as often as microphone B, or $H_0 : P_s = 50\%$) in most cases, does not imply that the null hypothesis is true [12]. However, we have been unable to find evidence that there is a consistent preference for a microphone type across subjects, even when a single source and fragment and a variation of microphone types and price ranges is considered.

The multiple stimuli test is substantially less time consuming, but in fact neither are too demanding in terms of time and effort for this number of different samples. It is therefore the opinion of the author that it is advisable for this kind of tests to include doubles of each microphone pair (not necessarily the exact same fragments) or, in the case of the multiple stimuli test, to simply do the test twice (again possibly with different programme content). This would increase performance and reliability as it allows for collection of more data per subject, and more information on intra-subject reliability.

Additional data and audio/video material of the recording session is available on www.brechtdeman.com.

6. ACKNOWLEDGEMENTS

The author wishes to thank Miss Maria Xifan Chen for donating her time and skill for the recordings, and everyone who participated in the listening tests.

7. REFERENCES

[1] S. P. Lipshitz and J. Vanderkooy, "The great debate: subjective evaluation," Journal of the Audio Engineering Society, Vol. 29, No. 7/8, 1981 July/August.

[2] Recommendation ITU-R BS.1770-2 *Algorithms to measure audio programme loudness and true-peak audio level*, 2011.

[3] W. Hoeg, L. Christensen and R. Walker, "Subjective assessment of audio quality - the means

and methods within the EBU," EBU Technical Review, 1997.

[4] S. Bech and N. Zacharov, *Perceptual audio evaluation*, John Wiley & Sons, 2007.

[5] Recommendation ITU-R BS.1534-1 *Method for the subjective assessment of intermediate quality level of coding systems*, 2001-2003.

[6] M. Zaunschirm, J. D. Reiss and A. Klapuri, "A sub-band approach to musical transient modification," Computer Music Journal, Vol. 36 (2), Summer 2012. Technical Report and Web page.

[7] S. Mansbridge, S. Finn and J. D. Reiss, "An Autonomous System for Multi-track Stereo Pan Positioning," 133rd AES Convention, San Francisco, 2012 October 26–29.

[8] S. Mansbridge, S. Finn, J. D. Reiss, "Implementation and Evaluation of Autonomous Multi-Track Fader Control," 132nd Audio Engineering Society Convention, Budapest, 2012 April 26–29.

[9] C. Uhle and J. D. Reiss, "Determined Source Separation for Microphone Recordings Using IIR Filters," 129th AES Convention, San Francisco, Nov. 4-7, 2010.

[10] S. Bech, "Listening tests on loudspeakers: a discussion of experimental procedures and evaluation of the response data," Proceedings from the AES 8th International Conference, Audio Eng. Soc., Washington, DC, 1990 May.

[11] L.E. Harris and K.R. Holland, "Using statistics to analyse listening test data: some sources and advice for non-statisticians," Proceedings of the Institute of Acoustics, 2009.

[12] L. Leventhal, "Type 1 and type 2 errors in the statistical analysis of listening tests," Journal of the Audio Engineering Society, Vol. 34, No. 6, 1986 June.