Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Musical Acoustics
Session 2pMU: Musical Preference, Perception, and Processing

2pMU7. Human ability of counting the number of instruments in polyphonic music

Fabian-Robert Stöter*, Michael Schoeffler, Bernd Edler and Jürgen Herre

***Corresponding author's address: International Audio Laboratories Erlangen, Am Wolfsmantel 33, Erlangen, 91058, Bavaria, Germany, fabian-robert.stoeter@audiolabs-erlangen.de**

There are indications that humans are only able to correctly count up to three voices in polyphonic music pieces of homogeneous timbre, where each voice is played by the same instrument. A more general case, where voices are played by instruments of inhomogeneous timbre, has not been fully addressed so far. In order to approach this question we conducted a listening experiment with 62 participants to find out whether both scenarios - instrumentation by inhomogeneous or homogeneous timbre - share the same outcome. This paper describes the design of the experiment including an analysis of the results, which show that both scenarios are related. Furthermore, a detailed analysis of the error rates in correctly counting the number of instruments reveals that there are significant differences between non-musician and musician listeners, in particular regarding the upper auditory limit of the number of correctly counted instruments. Based on these results, models for the perception of instruments in auditory streams can be developed.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4799609]
Received 22 Jan 2013; published 2 Jun 2013

INTRODUCTION

Decomposing music into its original audio sources can be a challenging task. Source separation methods can be analyzed how well they perform mathematically, but a human versus machine comparison is cumbersome because measurement of the human separation performance is problematic. One can easily evaluate if humans can detect the number of sources in a mixture of several sources. However, there are indications that even for this task, humans tend to fail if more than three sources are present at the same time [1]. Therefore, we want to take a first step by designing an experiment where we focus on polyphonic music of inhomogeneous timbre, where the question is: What is the number of instruments humans can estimate correctly? Several results are addressed in this paper including a possible upper limit of the number of perceived instruments, but also if one can see significant differences in performance of musicians compared to non-musicians. Such results can be used in auditory modeling or as a pre-processing step for source separation algorithms. This paper presents the results of an experiment, a detailed description of the stimuli and the statistical methods that were used.

RELATED WORK

The perception of concurrent sound sources has been analyzed on different scales so far. Bregman's and McAdams' [2] auditory stream theory can be seen as an analytical way of describing how sound events are perceived by the human auditory system. Unfortunately, it is difficult to model professionally produced music by auditory stream models because of its high complexity. As described by Wang [3] there exist several algorithms to estimate the actual number of sources. However, none of them is motivated to model the perceived number of musical sources. Kashino et. al [4] addresses the questions for concurrent speakers in a "cocktail party" like environment and found an upper limit of three voices humans can perceive. When the focus shifts to musical instruments as sources, research has to take concepts from musicology into account. Huron [1] was the first who addressed this question in 1989 at a musically meaningful level. Huron asked for the number of voices within a piece of music, where by voices in musicology one can define it as a line of sound or note events (See [5] for further definitions). Huron determined by experimental results that the number of correctly identified voices is up to three. Later in 1996, Reuter [6] has analyzed how combinations of different instruments are perceived which sets the focus on different instrumentation and not on denumerability.

EXPERIMENT

For the purpose of gaining more knowledge in understanding the human perception of multiple present instruments, an experiment was conducted. Huron selected voices from piano pieces only. We wanted to address the more general case where voices are played by different instruments. Therefore we used a method between musicology and auditory stream analysis to address this question. As we set our focus mainly on comparison between musicians and non-musicians, our experiment was designed so that it respects the fact that the latter have only limited musical background.

Although it might be interesting to have direct comparison with Huron's experiment, we agree that expanding the methods to an inhomogeneous timbre case is error prone. One reason is that there is reasonable doubt about the non-musicians understandings in terms of how a voice is defined. This is why we choose a trade-off with a more simplified experiment where we asked for the number of instruments instead of voices. Also whereas Huron [1] excluded subjects from his experiment because of their lower performance, we compared the results of both groups.

Stimuli

The selection of music items is crucial for our experiment setup. Usually music recordings have no ground truth metadata available to determine the actual number of instruments. Using annotated music like that from the RWC database [7] fulfills this requirement but lacks the possibility to remix, attenuate or suppress specific sources. This is important so that the experiment consists of equally grouped stimuli. Instead of the original RWC recordings, the annotated MIDI data itself was used as prototypes for the stimuli.

To make the counting task less ambiguous for the subjects, the instrumentation needs to be mostly constant during the music piece. Therefore we calculated an "instrumental stationarity" measure. The annotated MIDI files from [7] were converted into piano roll representations for each instrument channel. This representation was then converted into a binary *instrumentation activity matrix* $\mathbf{I}_{AM}[\underline{\mathbf{k}}_1|\underline{\mathbf{k}}_2|...|\underline{\mathbf{k}}_N] \in \{0,1\}$, where at each discrete time instant i a vector $\underline{\mathbf{k}}_i$ indicates which instruments are active. The aim is then to select frames of length N which are stationary by means of changes in instrumentation and activity. To get many items with a high instruments count, the maximum number of instruments within a frame was stored in a binary mask $\underline{\mathbf{k}}_{max}$ which was compared with all $\underline{\mathbf{k}}_{i=1...N}$ so that $(|\underline{\mathbf{k}}_i \oplus \underline{\mathbf{k}}_{max}| \leq 1) \vee (\underline{\mathbf{k}}_i = \underline{0})$. The resulting binary vector was smoothed with an averaging kernel of size N. By peak picking we got a list of possible candidates which contained a high stationarity in instrumentation. Further the RWC files were filtered a priori to exclude items dominated by electronic instruments or singing voice. Table 1 presents the selected 12 items representing pairs of one to six simultaneously present instruments. Each item is around seven seconds long. By cutting at note offsets we varied the lengths of the items to make it semantically more meaningful. Six items (notated as RM-C***) belong to the classical western music genre whereas the other items are of mixed genre.

The MIDI files were humanized randomly and rendered in a professional sequencer software utilizing state-of-the-art commercial sampling products. The rendered files were processed with convolutive reverb to better match the original recordings. Additionally equalization were applied to take loudness measures according to EBU R128 recommendations into account. To avoid spatial cues every item was rendered to mono at 16 bit/44.1 kHz.

			Instrument Name																		
RWC-MDB ID	Start [s]	Dur. [s]	Piano	Acoustic Guitar	Electric Guitar	Contrabass (pizz.)	Electric Bass	Violin	Viola	Violoncello	Contrabass	Trumpet	Trombone	French Horn	Tenor Sax	Oboe	Bassoon	Clarinet	Flute	Σ
J021**	46.5	6.6	×			×						×								3
C001	0.0	9.0															×			1
G047	35.3	8.3								×										
C016	0.9	7.6						×		×										2
G068	132.4	6.6						×											×	
C018	240.4	5.4	×					×						×						3
G046	0.3	7.9	×							×	×									
C013	5.6	6.0						×	×	×									×	4
G036	0.0	6.5	×	×			×	×												
C012	112.0	6.0						×	×	×	×								×	5
G037	67.1	7.0	×	×		×									×				×	
C001	147.8	6.0						×			×			×		×	×	×		6
G028	17.5	6.5	×		×		×					×	×						×	

TABLE 1: Selected items from the RWC Music Database [7]. Item *J021*** is used as training item.

Methods and Participants

The experiment was attended by 62 participants, where half of them regularly play a musical instrument. They were asked to count how many different instruments they can hear. 12 items from the test set (Table 1) were played back in random order. The experiment was presented by a user interface depicted in Figure 1. Except for the training item, every subject could play back each stimulus up to three times. Additionally they were asked to estimate how certain they were in their decision (ranged from *uncertain* to *very certain*). Instead of a slider

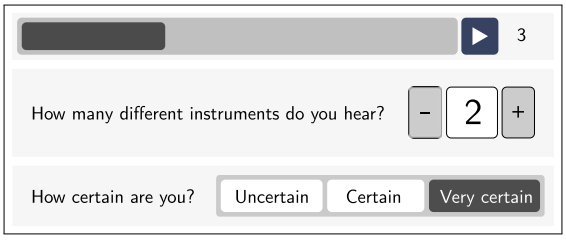

FIGURE 1: Experiment User Interface

UI-element, the interface only features plus and minus buttons so that the subjects were not biased about the maximum number of instruments. Item *J021*** had been selected as training item and was presented to the subjects during the introduction phase to make them familiar with the user interface. This trial also unveiled the number and name of the instruments within that piece. After they had read the introduction page, the subjects were asked to adjust the volume during the training example to their preference and leave the volume at that level for the duration of the experiment. The stimuli were presented on BEYERDYNAMICS DT770 headphones connected to a RME BABYFACE. The complete test took about 20 minutes on average for every participant.

RESULTS

The independent variable $I(i)$ is the number of instruments of one music item i where in this case $I(i) \in \{1,2,...,6\}$. $R(i,s)$ is defined as the number of instruments that are perceived and counted by subject s for music item i. The dependent variable is the derived from the main subject response as $\Delta(i,s) = R(i,s) - I(i)$ and can also be transformed into a binary scale:

$$E(i,s) = \begin{cases} 0 & \text{if } |\Delta| = 0 \\ 1 & \text{if } |\Delta| > 0 \end{cases}.$$

As we want to address a possible upper bound for the number of instruments that humans are able to correctly count. The primary statistical null hypothesis (H_1) is stated in that the means of Δ and E[1], grouped by the number of instruments, do not differ significantly. As we also want to test the between-groups performance of musicians versus non-musicians, we introduce another dependent variable $M(s) \in \{0,1\}$ of binary scale. This is stated in a secondary null hypothesis (H_2) where the means of Δ and E are not significantly different between musicians and non-musicians.

The outcome of the experiment is shown in Table 2. No subjects were screened from the results, although there are two cases where no valid response had been made. Results are grouped by items of I instruments.

In general participants tended to perform worse for items with more than two instruments. The probability of correctly counting one instrument was 90.0% whereas only one person out of

[1]The fact that E is dichotomous will lead to a mean value that equals to a probability of a binary distribution.

Responses		Stimuli of RWC-MDB items sorted by number of instruments												
		$I=1$		$I=2$		$I=3$		$I=4$		$I=5$		$I=6$		
Count	Subject Group	C001-1	G047	C016	G068	C018	G046	C013	G036	C012	G037	C001-2	G028	n
$R=0$	$M=0$	0	0	0	0	0	0	1	0	0	0	0	0	1
	$M=1$	0	1	0	0	0	0	0	0	0	0	0	0	1
$R=1$	$M=0$	25	30	7	6	0	0	1	1	2	0	1	0	73
	$M=1$	26	30	7	1	0	0	0	0	0	0	0	0	64
$R=2$	$M=0$	6	1	21	25	21	17	14	5	11	7	9	9	146
	$M=1$	5	0	18	29	6	4	3	3	3	0	3	0	74
$R=3$	$M=0$	0	0	2	0	8	14	11	20	11	17	13	13	109
	$M=1$	0	0	4	1	18	27	19	14	10	11	16	8	128
$R=4$	$M=0$	0	0	1	0	2	0	4	4	7	5	6	8	37
	$M=1$	0	0	2	0	7	0	9	13	15	18	5	22	91
$R=5$	$M=0$	0	0	0	0	0	0	0	1	0	2	2	1	6
	$M=1$	0	0	0	0	0	0	0	1	2	2	6	1	12
$R=6$	$M=0$	0	0	0	0	0	0	0	0	0	0	0	0	0
	$M=1$	0	0	0	0	0	0	0	0	1	0	1	0	2
	Σ	62 participants, 744 responses												
Error Probability	$M=0$	0.19	0.03	0.32	0.19	0.74	0.55	0.87	0.87	1.00	0.94	1.00	1.00	0.64
	$M=1$	0.16	0.03	0.42	0.06	0.42	0.13	0.71	0.58	0.94	0.94	0.97	1.00	0.53
	Σ	0.10		0.25		0.46		0.76		0.95		0.99		0.59
Mean($\Delta = I - R$)	$M=0$	-0.19	-0.03	0.10	0.19	0.61	0.55	1.48	1.03	2.26	1.94	3.03	2.97	1.16
	$M=1$	-0.16	0.03	-0.03	0.00	-0.03	0.13	0.81	0.61	1.39	1.29	2.45	2.23	0.73
	Σ	-0.09		0.06		0.31		0.98		1.72		2.67		0.94

TABLE 2: Experiment results by subjects. Gray background indicates correct responses.

62 gave a correct response for an item with six instruments. In some cases the number of instruments does not correspond to the number of voices for every item. Items where an instrument plays more than one voice and voices which are played by more than one instrument. However most of the chosen instruments are monophonic so in our case this occured only for items where piano or guitar is present. Also we made sure that the number of total voices did not exceed the maximum number of instruments in that item. Voices being played by more than one instrument are called unisono, which is present in G068, that showed surprisingly good results. The overall results of the means of Δ and E are plotted in Figure 2.

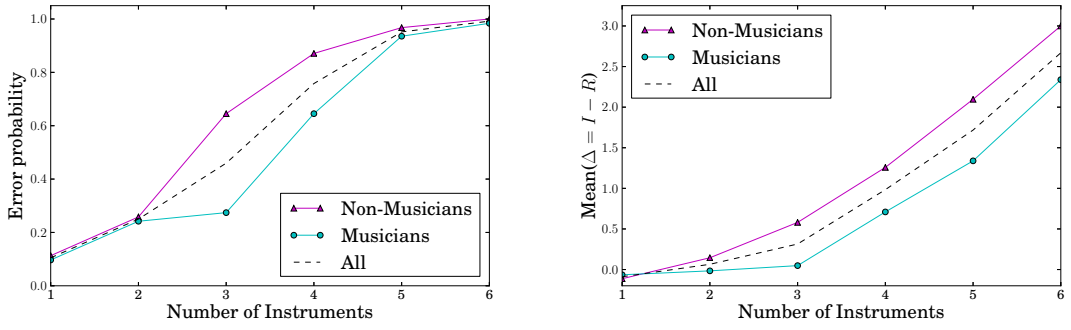

FIGURE 2: Error probability (left) and Mean of $\Delta = I - R$ (right) categorized by the number of instruments.

Underestimation

We confirm [1] that the most common error is the underestimation of one instrument, although this accounts only for 43 % of the responses in our experiment. Only in one case Δ is negative (overestimation) which is item C016, a "Clarinet Quintet in A major by Wolfgang Amadeus Mozart (K.581. 1st mvmt)" where we have excluded the solo clarinet part and two strings. Still, the remaining sound seems to be so similar to that of a quartet that musicians tended to hear "phantom" instruments.

Self-Evaluation

Figure 3 shows the results of the subjects certainty grouped by instrument count. Although the rate of "very certain" responses drops down to 11.3% for items with six instruments the rate of "certain" responses is still as high as 43.5%. When we take Δ into account we find a significant linear correlation between Δ and certainty where 0 is uncertain and 2 is very certain (Pearson's $r = -0.227$ at the $p = 0.05$ level).

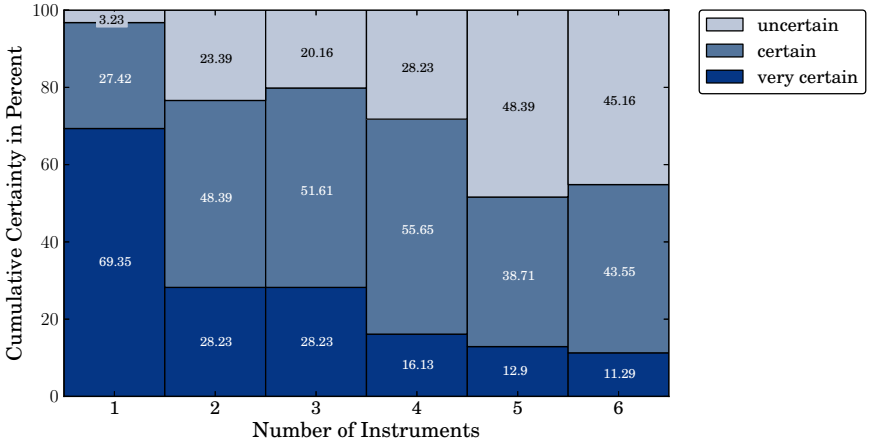

FIGURE 3: Responses for certainty of subjects by number of instruments

Main Effects

Rejecting the null hypotheses (H_1 and H_2) requires further statistical methods. One reason is that the dependent variables Δ and E have different scales. We will show that both variables qualify to reject our hypotheses. As E is dichotomous we focus on Δ first which can be classified as an interval scaled variable. To show differences between means of two or more groups, usually One-Way-ANOVA tests are applied. ANOVA tests expect independent normally distributed variables and homogeneity of the variances in each group. However both the Kolmogorov–Smirnov test of normal distribution and Levene's test to determine the homogeneity of group variances fail. In such cases variables scaled like Δ could be transformed so that the boundaries are straightened out. A typically used $arcsin(\sqrt{\Delta})$ transformation was applied to Δ resulting in slightly higher p values but still not statistically significant. Although ANOVA is known to be robust enough to run the tests against non-normal distributed cases and unequal variances, the significance levels of the results are doubtful. Therefore we choose to run a non-parametric test. The Kruskal–Wallis test can be applied even if the data is not normally distributed. However it has to be run on a slightly modified hypothesis which compares the medians of groups instead of the means. The Kruskal–Wallis test allows to reject both modified hypotheses (asymptotic $p = 0.000$, $\chi^2 = 499636$, $df = 5$). As mentioned by [8] it may be crucial for the Kruskal–Wallis test to be run on groups with differently shaped distributions. This may

Parameter	B	Std. Error	Wald χ^2	df	Sig.
(Intercept)	-4.456	1.0067	19.590	1	.000
$M = 0$	-.910	.2147	17.977	1	.000
$M = 1$	0	.	.	.	
$I = 1$	7.136	1.0507	46.131	1	.000
$I = 2$	6.061	1.0291	34.691	1	.000
$I = 3$	5.081	1.0228	24.677	1	.000
$I = 4$	3.715	1.0274	13.076	1	.000
$I = 5$	1.841	1.0892	2.856	1	.091
$I = 6$	0	.	.	.	.

TABLE 3: Parameter Estimates for GLM model. Missing values are represented by the Intercept.

not be assured in the case of Δ as the skewness highly increases for $I \geq 3$. Although the significance of the test has to be treated with caution, the median is a good indicator of a possible upper limit. The overall medians for all groups are plotted in Figure 4 where one can see that three instruments are a possible upper bound.

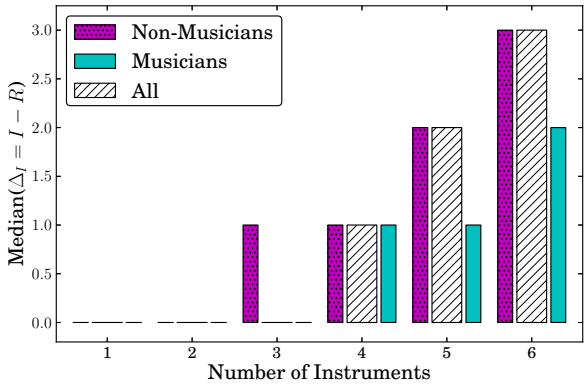

FIGURE 4: Median of Δ categorized by the number of instruments

The reason why an ANOVA is not recommended to be run on E is because it is a categorical variable. Like described in [9] this leads to problems which can be avoided by using a binary regression model that turns the mean of E into a an binomial distributed probability. Whereas in standard linear models like ANOVA the output variable will be modeled by a linear equation, in a so called *binary logit regression* the output will be modeled by *log linear* values. By including the main factors I and M we set up a *Generalized Linear Model* (GLM)

$$logit(E) = \text{Intercept} + x_1\text{I} + x_2\text{M}. \tag{1}$$

A test of the main effects is statistically significant ($\chi^2 = 437418$, $p < 0.000$, $df = 6$) so that both null hypotheses (H_1 and H_2) can be rejected. The significance of both effects is shown in Table 4 whereas parameter estimates and Wald values of the calculated model are shown in Table 3. The intercept represents a constant calculated offset within the regression.

The tests we introduced in the last section show that there is a significant difference in the error probability for groups of different instrumentation but also for musicians versus non-musicians. A pairwise comparison test based on the mean differences reveals where these differences are located. Regarding the error probability of different instrument counts, the pairwise comparison test reveals that nearly all groups show significant mean differences between each other, which was the expected result. However, by calculation using the logit GLM

Source	Wald χ^2	df	p
Intercept	22.643	1	0.000
I	185.359	5	0.000
M	17.977	1	0.000

TABLE 4: Tests of Model Effects of dependent variable E

Source	Wald χ^2	df	p
Intercept	12.434	1	0.000
I	22.742	1	0.000
M	22.742	1	0.000
M×I	0.184	1	0.668

TABLE 5: Tests of Model Effects of dependent variable E for a subset of items with $I \in \{3,4\}$

model shown in equation 1 we found that there are two groups of items of five and six instruments (mean difference 0.04, std. error = 0.019, $df = 1$, $p = 0.055$) that did not show any significant difference. For both groups the error probability is close to 100%.

A gap of one instrument between musicians and non-musicians

To investigate the difference in performance between musicians and non-musicians a pairwise comparison between those two groups was run. Overall musicians perform about 20% better throughout the test (mean difference = 0.18, std. error = 0.0044, $df = 1$, $p = 0.000$). We do not know what caused these differences as the level of professionalism had not been surveyed. Also 37 % of the musicians additionally had experience in audio engineering due to their profession. Further to look at possible interaction effects between the number of instruments and the groups of musicians and non-musicians we adapted our logit equation to

$$logit(E) = \text{Intercept} + x_1\text{I} + x_2\text{M} + x_3\text{M} \times \text{I}. \tag{2}$$

We then reran the GLM analysis selecting only items of three and four instruments. This avoids quasi-complete separation in the logit regression model which is caused by low variances in the error probability for items of $I \in \{1,2,5,6\}$. The model effects of the subset of items are shown in Table 5.

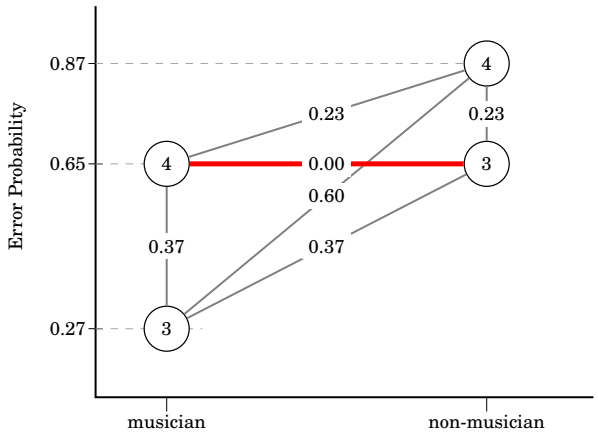

FIGURE 5: Pairwise comparison between interaction of Musician/Non-Musician and the number of instruments (labeled in nodes). The costs between nodes indicates the mean differences between groups. The red/bold line indicates that is there is no significant difference at the $p = 0.005$ level

The interaction of Musician×Instruments is not significant on a $p = 0.05$ level in general and a pairwise comparison test reveals two groups of equal probability. The pairwise comparisons are depicted in Figure 5. The red vertex indicates there is no significant difference in the error probability for the group of musicians in items with four instruments compared to

non-musicians in items of three instruments. Therefore a gap in the error probability of one instrument between those two groups becomes apparent.

Conclusion

This paper shows that counting the number of instruments in a music item is a difficult task for humans. A new experiment with 62 participants was conducted to address the question of how many instruments one can estimate correctly. In this experiment, the focus was set on stimuli of inhomogeneous timbre and also mixed genre. By comparing musicians to non-musicians, we revealed that there is a significant difference in performance. Particularly this gap is most prominent for items of three and four instruments. Furthermore, for all stimuli (ranging from one to six instruments) we see that musicians performed about 20% better than non-musicians. The experiment shows an assumed upper limit for items with more than three instruments. Our results are closely related to a previous experiment which focused on voices instead of instruments. This work can be used as a basis to define target functions for related research in auditory modeling. Future work could address different groups of items to reveal if certain instrumentations or compositions influence the results.

References

[1] D. Huron, “Voice denumerability in polyphonic music of homogeneous timbres”, Music Perception 361–382 (1989).

[2] S. McAdams and A. Bregman, “Hearing musical streams”, Computer Music Journal 26–60 (1979).

[3] D. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, 68–69 (IEEE Press) (2006).

[4] M. Kashino and T. Hirahara, “How many concurrent talkers can we hear out”, in *Proc. ASJ*, 3–3 (1995).

[5] E. Cambouropoulos, “Voice and stream: Perceptual and computational modeling of voice separation”, Music Perception **26**, 75–94 (2008).

[6] C. Reuter, *Die auditive Diskrimination von Orchesterinstrumenten* (1996).

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical, and jazz music databases”, in *Proc. ISMIR*, volume 2, 287–288 (2002).

[8] M. Fagerland and L. Sandvik, “The wilcoxon–mann–whitney test under scrutiny”, Statistics in Medicine **28**, 1487–1497 (2009).

[9] T. Jaeger, “Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models”, volume 59, 434–446 (Elsevier) (2008).