DURMS
Durham Undergraduate
Research in Music & Science

Categorisation of Extremely Brief Auditory Stimuli: Speech and Music Recognition

Kiko Keighery, Alessandra Douglas, Archie Jury, Alejandra Malave
Durham University

ABSTRACT

The ability to recognise very brief auditory stimuli is fundamental to everyday, rapid decision making. Given this, the present study investigated the accuracy of differentiation between human speech and instrumental music at extremely short durations, replicating the study conducted by Bigand et al. (2011). Furthermore, expansions to the original study were made by also exploring the impacts of musicality and music genre. Participants were asked to identify music and speech from extremely short auditory stimuli through an online survey. Several new findings were reported. 1) The success rate was very high on average, even at 20 ms. In line with the replicated study, the time interval between 20 and 30 ms produced the most significant improvement in categorisation. 2) Musical expertise does not make a significant impact on categorising audio of these durations. 3) There was no significant difference in the categorisation of classical or electronic music. Some of the findings challenge previous literature on the significance of musical expertise and instrument/genre on categorisation and raise questions as to whether individuals are consciously identifying the sound source or differentiating by more general timbral groups. Useful recommendations for future research have been provided on procedure and direction to confirm the present findings.

1. INTRODUCTION

Auditory information is vital for most people in recognising emotions, events, and objects in their surrounding environment. Oftentimes this information must be gathered through very brief or minimal stimuli in order to make important decisions rapidly, such as when driving a vehicle. Studies using ERP scans have shown that vocal stimuli elicit brain activity localised in different locations than those elicited by man-made auditory objects (Murray et al., 2006) and every-day life sounds (Charest et al., 2009). As these stimuli are processed in different ways, there may be reason to suspect that this affects the conscious ability to differentiate between these and other types of sound, like types of music. Given the following studies, it appears this is possible even without typically present contextual elements such as key, harmonies, rhythm, and language.

Krumhansl (2010) studied participants' ability to recall details of popular songs from short clips (300 and 400 ms) from memory. They found that participants were able to consciously discern both artist and title accurately from 400 ms clips on over 25% of the trials and, even if a clip was not identified, identify emotional content, style, and (in some cases) decade of release. Gjerdingen and Perrott (2008) investigated how quickly participants could recognise genres of music and similarly found that participants were unexpectedly successful in identifying genres for clips of 250 ms to 475 ms in duration. They also discovered that instrumental clips produced a higher success rate than vocal styles at every time interval, suggesting this is due to recognition of the distinct timbre typically associated with each genre.

Evidently, the focus is divided within the previous literature across a numerous range of musical styles and genres which, despite producing a vast amount of data, does not pinpoint where the boundaries of recognition lie. It is therefore of curiosity to investigate how adept we are at differentiating between human speech and instrumental music. Bigand et al. (2011) begin to narrow down this area in the following study, which the present study will aim to replicate.

Original study. The replicated study is titled ‘Categorization of Extremely Brief Auditory Stimuli: Domain-Specific or Domain-General Processes?’ by Emmanuel Bigand, Charles Delbé, Yannick Gérard, and Barbara Tillmann (2011). It investigates the minimum amount of auditory stimulation that is required to differentiate between speech, environment sounds, and music, as well as the effect of two types of amplitude normalisation: peak-level and RMS normalisation. The study consisted of a sample of 37 participants with an age range of 18 to 25 (psychology students with no formal music training and no hearing difficulties). Eighteen participants completed the task in a peak-normalisation condition, and the other 19 in an RMS-normalisation condition. The musical excerpts were taken from a collection of classical and romantic symphonies. The speech stimuli were taken from radio channels of single voices speaking, with a mixture of both male and female voices. The environmental sounds were selected from audio CDs, specifically ones with a high probability of occurrence in

an everyday-life environment. Twenty excerpts for each category of sounds (3) were used in each duration condition (5), generating 300 (20×3×5) stimuli. Each participant heard the 300 stimuli played in a random order, gradually moving up in duration starting with the ones of 20 ms, then 30 ms, 50 ms, 100 ms and finally up to the 200 ms.

Bigand et al. (2011) found that accuracy increased with duration, with an increase of just 10 ms between 20 ms and 30 ms being sufficient to significantly improve performance in all sound categories. The other time increments had a smaller impact, but still attained statistical significance. Recognition for both voice and music categories was above chance level at 20 ms for the peak-level normalisation and 30 ms for RMS normalisation, as was environment sounds at 30 ms for peak-level and 50 ms for RMS normalisation. These findings therefore demonstrated an overall higher accuracy in the music and voice categories over the environment sounds category. Moreover, when RMS energy of the stimuli was equalised, voice recognition was most successful. All participants (100%) complained about the difficulty of the task despite showing that a score of 100% was still attainable by some. In contrast to Gjerdingen and Perrott's findings (2008), participants reported finding voice stimuli easier to detect, however, this voice superiority effect was only visible in the results in the RMS normalisation condition.

Modifications to the present study. Bigand et al.'s (2011) study contained three categories of auditory stimuli and longer durations. Environmental sounds will be removed in this replication study to focus on the differentiation between speech and music. The longest duration of auditory stimuli (200 ms) will also be removed due to the high success rate shown in the original study. Within the category of music, subgroups of 'classical' and 'electronic' music will be added to allow insight into whether the genre of music would also have any impact on the categorisation of the auditory stimuli. The different normalisation conditions will be removed to focus on the audio duration. For the purposes of increasing the chances of participants fully completing the study, the number of clips will be significantly reduced to keep the expected experiment duration below 10 minutes.

Hypotheses. We hypothesise that participants' success in categorising auditory stimuli will increase positively with the duration, in line with the findings of Bigand et al. (2011), and that 20 ms auditory stimuli will have a low identification success rate (Hypothesis I). We have no major expectations of musicianship making a significant difference in the ability to categorise the clips (Hypothesis II), particularly at the shortest duration of 20ms. However, previous studies have indicated that musical expertise does impact quality of auditory perception, including nuanced elements such as harmonic dissonance (Linnavalli et al., 2020). Dawson et al. (2017) found musical sophistication to have an impact on pitch discrimination in clips as short as 25ms, though there was no notable impact on duration discrimination. In light of this, we suspect musicians may have an advantage at the other time intervals. It is predicted that classical music stimuli will be slightly more difficult to differentiate from speech than electronic music due to the difference in timbre (Hypothesis III), though similar results are expected for both in accordance with Gjerdingen and Perrott (2008).

2. METHODS

Stimuli. The study contained speech stimuli and musical excerpts, totalling 40 stimuli. The speech stimuli were gathered from an online database called VoxCeleb (Nagrani et al., 2020). These stimuli were all extracted from interview videos uploaded to YouTube. The music stimuli were gathered from the Emotify dataset (Aljanaki et al., 2016). To avoid any other elements such as instrument or the gender of the speaker affecting the results of the study, the stimuli were selected with variety in mind. The 20 speech stimuli contained an equal number of female-sounding speech and male-sounding speech, whilst the musical stimuli contained a mixture of instruments including piano, strings, woodwind, and electronic sounds. Among the 20 music stimuli, 10 of these were categorised as 'classical' in the Emotify dataset, while the remaining 10 were categorised as 'electronic'. The stimuli were all edited in Audacity. The clips were normalised to maintain a standard peak amplitude across the entire set of stimuli and cut down to the relevant length (20ms, 30ms, 50ms, or 100ms). Each clip included a silent buffer of approximately 30 ms prior to the auditory stimuli to account for potential audio delays. These were imported into the survey as separate questions on Qualtrics for the participants to access remotely.

Ethics and collected data. Prior to starting the main content of the study, participants were asked to read through a briefing sheet and consent to their participation, before answering several demographic questions. Consent was fully informed, and participants were allowed to withdraw at any point during the study. The age and gender of each participant was collected as well as their level of musical expertise. This information was collected using the self-report one-item version of the Ollen Music Sophistication Index (OMSI) (Zhang and Schubert, 2019).

Design. Following the demographic questions, participants were asked to listen to and categorise each of the extremely short auditory stimuli as either speech or music. The study was made with a within-subjects design, therefore all participants listened to the same 40 auditory stimuli in the same order. It was possible to listen to the clips multiple times and there were no time constraints. The independent variable was the auditory stimuli duration, and the dependent variable was the accuracy of categorisation. Musical expertise and genre were investigated as secondary independent variables. The study began with the shortest duration stimuli and finished with the longest in the same way as the original study. Questions 1 to 10 contained the 20 ms clips, 11 to 20 contained 30 ms clips, 21 to 30 contained 50 ms clips and 31 to 40 contained the 100 ms clips. After the participants had answered to each of the stimuli the study was complete. Completion time was approximately 5 minutes for most participants.

Participants. Participants were recruited through volunteer sampling. An online link to the study was distributed through social media posts and messages to give access to a wide range of participants. A total of 65 people fully completed the survey, with no additional participants shown to have only partially completed it. The mean age of the sample was 35.2 years. Ages ranged from 18 to 63. The sample contained 28 males, 36 females, and 1 participant who preferred not to disclose this information. The musical backgrounds of the sample are as follows: 13 non musicians, 23 music-loving non musicians, 16 amateur musicians, 6 serious amateur musicians, 7 semi-professional musicians. Thirty-two participants used headphones while 33 did not.

3. RESULTS

*Analytic plan***.** Table 1 displays the values of the average score achieved by each group at each duration. To examine whether the duration of stimuli had a positive relationship with the participant's success rate (Hypothesis I), three *t*-tests were conducted between each time interval. A graphical representation of the means for each duration along with the 95% confidence intervals can be seen below in Figure 1. To examine the second hypothesis, two *t*-tests were conducted: one comparing the non-musician group to the semi-professional group at the 20 ms duration, and one comparing all non-musicians (non-musicians and music-loving non musicians) and all musicians (amateur, serious amateur and semi-professional) at all durations. A graphical representation of Table 1 can be seen below in Figure 2. A graphical representation of the overall performance for each musical skill group can also be seen below in Figure 3. Both figures contain the 95% confidence intervals. To examine the final hypothesis, a *t*-test was conducted to compare the average total correctly categorised classical music stimuli and electronic music stimuli. A graphical representation of this with the 95% confidence intervals can be seen in Figure 4.

Table 1. Mean number of correctly categorised auditory stimuli at each duration out of 10

Musicianship	20ms	30ms	50ms	100ms	Total
Non-musician	6.08	8.23	8.46	9.00	31.77
Music loving Non-musician	6.52	8.52	8.65	9.39	33.09
Amateur Musician	6.64	8.64	8.82	9.23	33.32
Semi-professional Musician	7.43	8.57	8.43	9.14	33.57
All	6.57	8.51	8.65	9.23	32.95

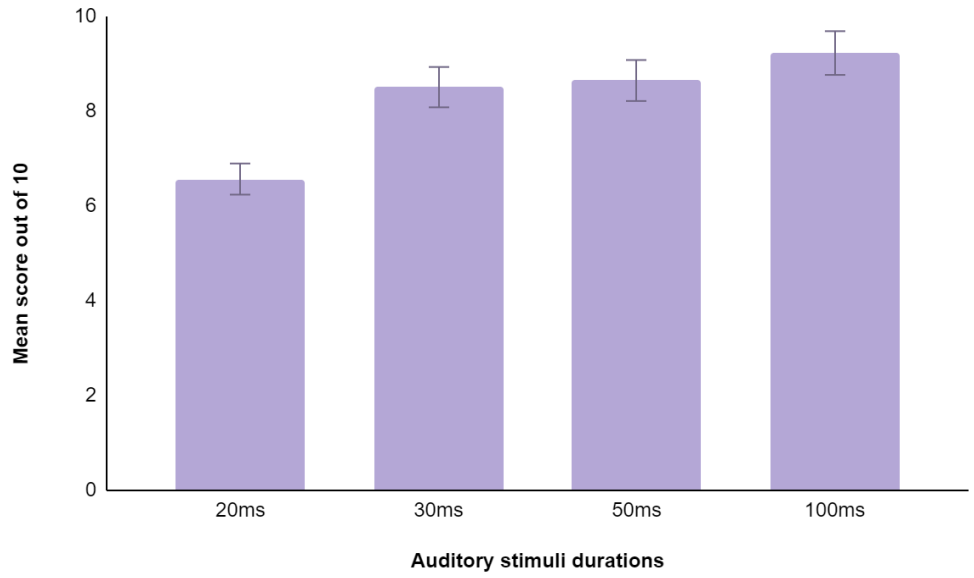

Figure 1. Mean number of correctly categorised auditory stimuli across all participants at each duration

There was a significant difference between the 20 ms condition (M = 6.57) and 30 ms condition (M = 8.51) on successful categorisation of auditory stimuli, $t(128) = -7.10$, $p < .001$. There was no significant difference between the 30 ms condition (M = 8.51) and 50 ms condition (M = 8.65) on successful categorisation of auditory stimuli, $t(127) = -0.53$, $p = .59$. There was a significant difference between the 50 ms condition (M = 8.65) and 100 ms condition (M = 9.23) on successful categorisation of auditory stimuli, $t(126) = -2.52$, $p = .01$.

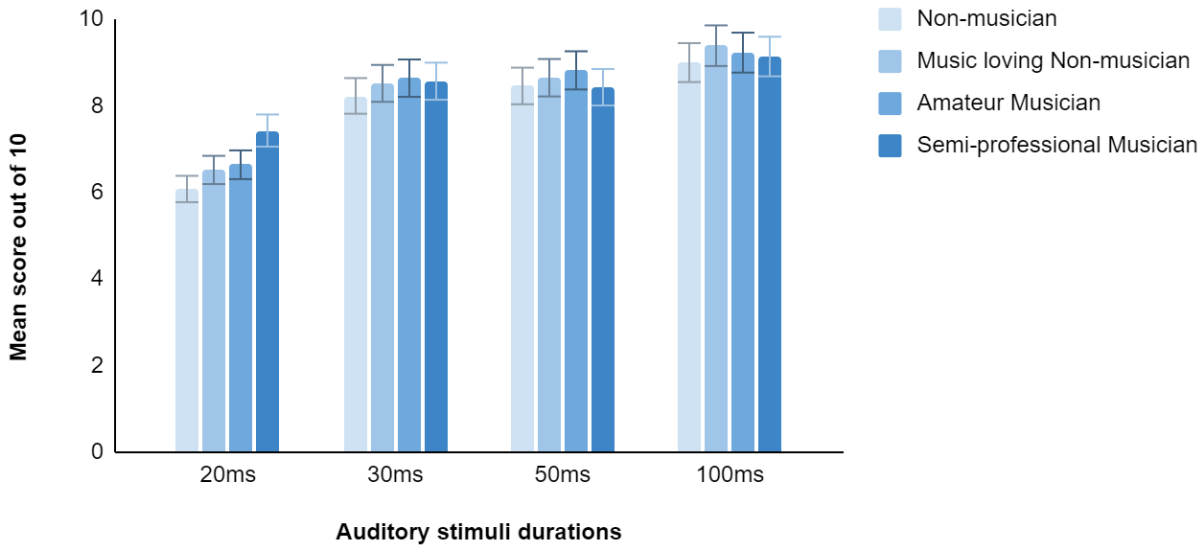

Figure 2. Mean number of correctly categorised auditory stimuli at each duration, separated by each musical skill group

There was a marginally significant difference between the non-musicians group (M = 6.08) and semi-professionals group (M = 7.43) on successful categorisation of auditory stimuli at the 20 ms duration, $t(15) = -2.06$, $p = .05$.

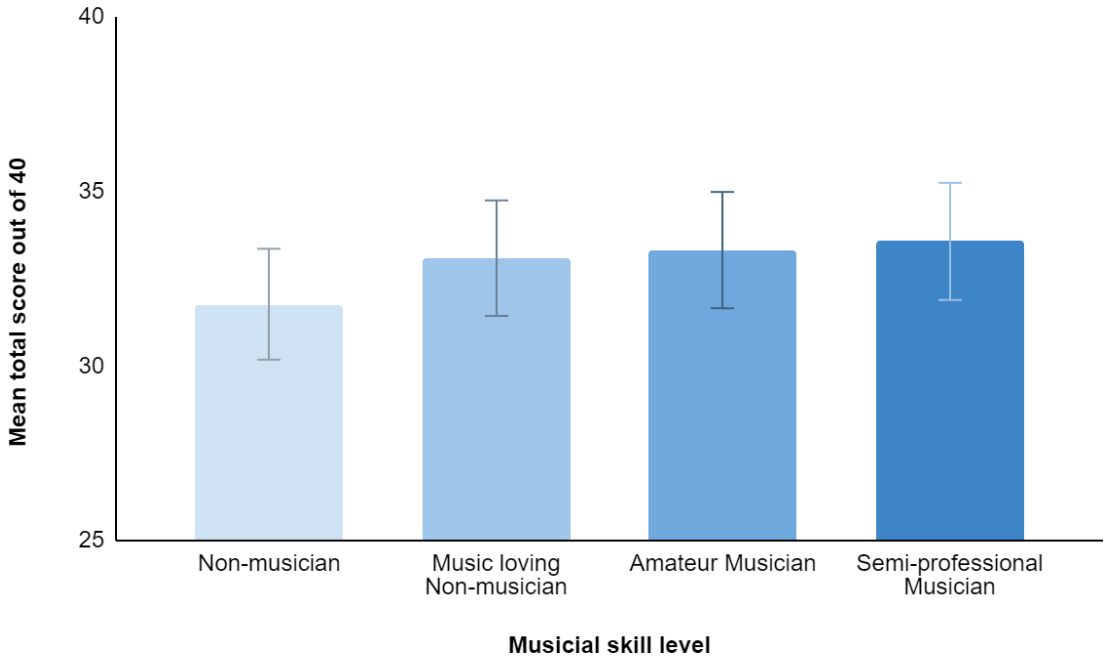

Figure 3. Mean number of correctly categorised auditory stimuli by each musical skill group

There was no significant difference between non-musicians (the combined total of both non-musicians and music-loving non-musicians) (M = 32.61) and musicians (the combined total of amateur musicians and semi-professional musicians) (M = 33.38) on successful categorisation of all the auditory stimuli, $t(57) = -0.66$, p = .50.

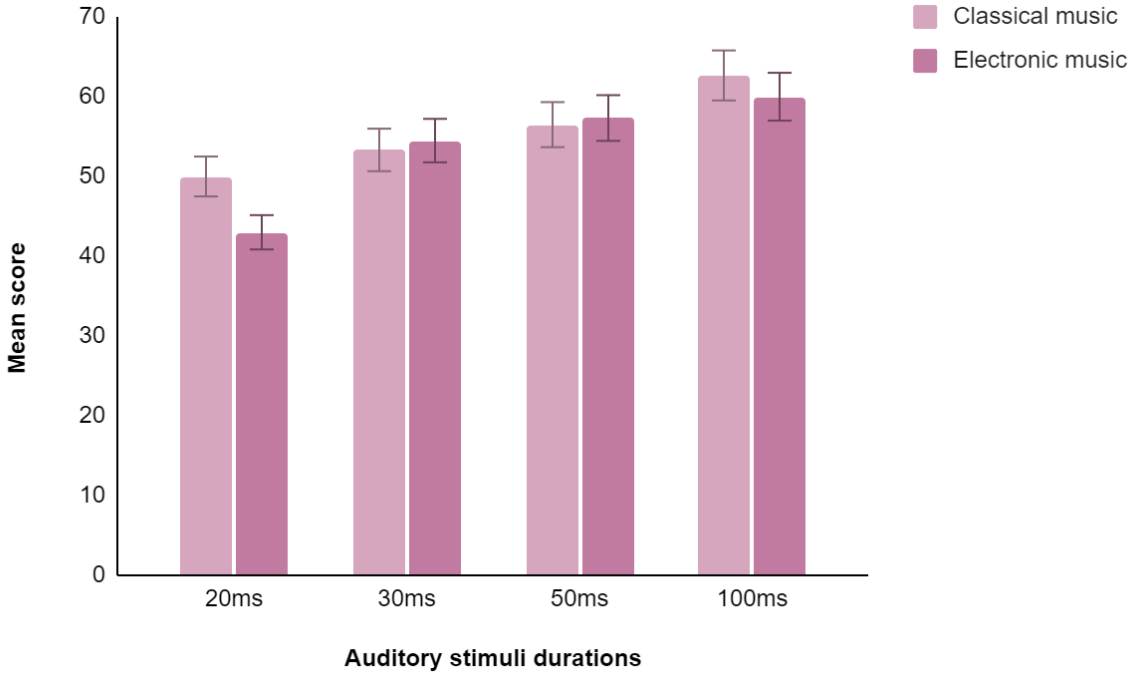

Figure 4. Mean number of correctly categorised musical stimuli at each duration, separated by music genre

There was no significant difference on the participants' success rate in correctly categorising music stimuli between classical music ($M = 55.63$) and electronic music ($M = 53.71$), $t(5) = 0.42$, $p = .69$.

4. DISCUSSION

This study aimed to investigate the shortest duration at which successful sound categorisation can occur. Additional investigations were made into musicianship and genre of music in relation to this, as they have not previously been studied in this specific context. It was hypothesised that auditory stimuli of 20 ms would cause difficulties in categorisation, with accuracy increasing positively with duration. The results supported the predicted correlation, but showed there was a high success rate, above the chance level of 50%, for all durations. Musicality and genre were hypothesised not to have a significant impact on categorisation accuracy. The results show there were no major statistically significant effects due to either condition.

Hypothesis I. In terms of categorisation accuracy, the present findings align with that of Bigand et al. (2011). Both studies found accuracy increased with duration, predominantly with statistical significance, with the time interval between 20 ms and 30 ms producing a drastic increase in accuracy with significant results. However, the 30 ms to 50 ms interval lacked a significant increase in the present study unlike the Bigand et al. (2011) study. It is possible the larger sample has emphasised the significance of the 20-30 ms interval and minimised that of the remaining intervals. Despite producing significant increases in accuracy at most of the duration intervals, the present study had a surprisingly high level of accuracy across all durations. From 30 ms upwards accuracy was consistently high on average, an improvement on the results from the original study likely impacted by the lack of a third sound category. Participants experienced some difficulties with the 20 ms stimuli, though all achieved a score above chance level despite these stimuli being first to appear in the study with no training questions to precede them, suggesting individuals may be able to identify stimuli of an even shorter duration. Similar to Bigand et al.'s (2011) study, optional feedback from participants following the completion of the study informed us that most of the participants perceived the study as a difficult task. Despite this perceived difficulty, our results indicate that participants were still able to categorise auditory stimuli better than they thought. This suggests a disconnect between the perceived personal ability and actual cognitive ability of the participants.

Hypothesis II. In line with the hypothesis, average overall scores of each musical skill group were quite close with no significant difference between any groups. This opposes past literature (Linnavalli et al., 2020; Dawson et al., 2017), therefore suggesting that identification of extremely short auditory stimuli is not necessarily impacted by musical expertise, though there were no fully professional musicians in the present study. Interestingly, there was a marginally significant difference between non-musicians and semi-professional musicians on average at the 20 ms duration, therefore it may be possible that musical expertise has an impact on recognition at durations below 30ms. This cannot be confirmed with the conditions investigated in the present study, which guarantees future research. However, it should be noted that the music loving non-musician group achieved a closer average score to the musician groups than the non-musician group. This may indicate the development of the human brain to better process auditory stimuli is not exclusively assisted by traditional musical training, but also by exposure to music in a non-academic setting.

Hypothesis III. As Gjerdingen and Perrott (2008) suggested, classical and electronic music stimuli produced relatively similar results of recognition accuracy. No significance was found in the overall scores, and no relationship or consistent correlation was found between the conditions. Due to each of the genres only constituting 25% of the stimuli, a larger dataset may find alternative results.

Strengths, limitations, and future directions. The present study found a surprisingly high accuracy on all durations and musicality levels despite the lack of a practice task), as well as a highly significant p-value between the 20 ms and 30 ms duration increments. It would be of interest to investigate auditory stimuli at durations smaller than 20 ms in future investigations to locate the limit of audio category recognition. It may be of value to also study smaller increments between 20 ms and 30 ms in order to identify the duration at which recognition is more successful, or if this is a gradual improvement between the two values.

The study also showed good engagement, with all participants who started the study completing it all the way through. Conducting the study online rather than in person enabled a larger sample than the original study (Bigand et al., 2011). Age, gender, musical expertise, and profession were not restricted in the present study, contributing to the larger sample. This decreases validity when comparing with the Bigand et al. (2011) study, however, it was found that age and gender made no impact on categorisation success. Participants aged 50+ years were amongst those who achieved both the highest and the lowest overall score of 39 and 23 (out of a

possible 40), respectively. The wider sample validates and extends further the findings of the original study because it was limited to undergraduate psychology students aged 18-25, which has the potential for bias.

As an online study, it was not feasible for participants to all have access to the same equipment or environment and so this must be considered as an extraneous variable. Participant feedback revealed that it was possible to tell between stereo (music stimuli) and mono (speech stimuli) when using headphones. In future online studies it is recommended to ensure all stimuli are mono. Furthermore, it was also not possible to control how many times participants could listen to each stimulus, due to potential technical issues which may require them to replay the audio multiple times. There was also potential for participants to receive help from others around them. However, despite not being conducted in controlled conditions, the present study still produced similar results to the original study suggesting the findings are robust enough to remain consistent outside of a lab environment. It may be of value to further investigate the effects of replaying the audio and the use of headphones of categorisation accuracy, though it is recommended to conduct these in person in future to ensure internal validity of results.

Data regarding hearing disabilities was collected, with six participants reporting to have some level of hearing impairment. No correlation between these and auditory categorisation success was found, but with such a small sample no significant conclusion can be made. It is worth noting that these participants with hearing impairments who used headphones achieved marginally higher scores. Previous research by Hiraga and Otsuka (2012) found that individuals with hearing difficulties can successfully differentiate timbres but not identify specific sources. It is likely that participants in this study were making guesses based on timbre in this way rather than consciously recognising each stimulus specifically as music or speech. In future research it may be useful to have an 'unsure' option to ascertain how often participants are guessing or consciously identifying the stimuli.

Conclusions and implications. The present research expanded on previous literature on categorisation of brief auditory stimuli with additional investigation into musicality and genre within this. Similar results were found to the replicated study (Bigand et al., 2011) with modifications that lead to useful recommendations being put forward to improve future investigations. Firstly, time increments should be adjusted to identify definitive limits to categorisation accuracy, specifically shorter intervals between 20 and 30ms, and perhaps increments below 20ms. Secondly, research focussing on the identification of a wider range of genres at extremely short durations may be of interest in timbre recognition. Thirdly, similar studies are recommended to be conducted in person with standardised equipment in order to achieve results of increased validity. Finally, investigation into hearing difficulties must be executed independently due to the variation of impairments. Including an 'unsure' option as one of the possible responses may help differentiate between recognition of timbral variability and specific identification of auditory stimuli. Adhering to the above suggestions in future studies may help determine whether there are any processing advantages such as the voice superiority effect beyond lab environments.

REFERENCES

Agus, T. R., Thorpe, S. J., Suied, C., & Pressnitzer, D. (2010). Characteristics of human voice processing. *IEEE International Symposium on Circuits and Systems (ISCAS),* 509-512.

Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2016). Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management, 52*(1), 115-128.

Bigand, E., Delbé, C., Gérard, Y., & Tillmann, B. (2011). Categorization of extremely brief auditory stimuli: domain-specific or domain-general processes? *PLoS ONE*, *6*(10), e27024.

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., & Latinus, M., et al. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, *10*(1), 127.

Dawson, C., Aalto, D., Šimko, J., Vainio, M., & Tervaniemi, M. (2017). Musical sophistication and the effect of complexity on auditory discrimination in finnish speakers. *Frontiers in Neuroscience, 11*, 213.

Gjerdingen, R. O., & Perrott, D. (2008). Scanning the dial: the rapid recognition of music genres. *Journal of New Music Research, 37*(2), 93-100.

Hiraga, R., & Otsuka, K. (2012). On the recognition of timbre - a first step toward understanding how hearing-impaired people perceive timbre. *IEEE International Conference on Systems, Man, and Cybernetics (SMC),* 2103-2108.

Krumhansl, C. (2010). Plink: 'thin slices' of music. *Music Perception, 27*(5), 337-354.

Linnavalli, T., Ojala, J., Haveri, L., Putkinen, V., Kostilainen, Sirke Seppänen, S., & Tervaniemi, M. (2020). Musical expertise facilitates dissonance detection on behavioral, not on early sensory level. *Music Perception*, *38*(1), 78–98.

Murray, M. M., Camen, C., Gonzalez Andino, S. L., Bovet, P., & Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *The Journal of Neuroscience, 26*(4), 1293-1302.

Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: large-scale speaker verification in the wild. *Computer Speech & Language,* 60, 101027.