The language of sounds unheard: Exploring sensory semantic knowledge in large language models

Kai Siedenburg[*a], Charalampos Saitis[*b]

[a]Dept. of Medical Physics and Acoustics, University of Oldenburg

[b]Centre for Digital Music, Queen Mary University of London

[*]equal contribution

April 27, 2023

Abstract

Semantic dimensions of sound have been playing a central role in understanding the nature of auditory sensory experience as well as the broader relation between perception, language, and meaning. Accordingly, and given the recent proliferation of large language models (LLMs), here we asked whether such models exhibit an organisation of perceptual semantics similar to those observed in humans. Specifically, we prompted ChatGPT, a chatbot based on a state-of-the-art LLM, to rate musical instrument sounds on a set of 20 semantic scales. We elicited multiple responses in separate chats, analogous to having multiple human raters. ChatGPT generated semantic profiles that only partially correlated with human ratings, yet showed robust agreement along well-known psychophysical dimensions of musical sounds such as brightness (*bright–dark*) and pitch height (*deep–high*). Exploratory factor analysis suggested the same dimensionality but different spatial configuration of a latent factor space between the chatbot and human ratings. Unexpectedly, the chatbot showed degrees of internal variability that were comparable in magnitude to that of human ratings. Our work highlights the potential of LLMs to capture salient dimensions of human sensory experience.

1 Introduction

It has recently become apparent that language processing based on large language models (LLMs) is not only here to stay, but also about to change research and society. This insight dawned most strikingly with the surfacing of chatbot *ChatGPT*[1] (CGPT, *Chat Generative Pre-trained Transformer*), based on the LLM *GPT3* from by the company OpenAI[2]. CGPT mimics a human conversationalist and seemingly covers any type of topic. A major line of critique of CGPT yet concerns its relation to facts for it often provides answers that sound plausible but are incorrect. For perception research, an important question concerns the extent to which LLMs faithfully recover dimensions of human experience. Specifically, the extent to which the perceptual semantics of LLMs map onto human perceptual semantics remains unclear. Marjieh et al. (2023) observed that GPT3 structurally aligns with human auditory experience, as probed with dissimilarity judgments of consonants, timbre, pitch, and loudness. Here, we wish to explore the extent to which the chatbot CGPT aligns with human judgments of fine grained semantic dimensions of sound.

Although auditory sensory experience is rich, humans lack a rich generic vocabulary for sound. In fact, most Western languages have relatively few generic adjectives to describe auditory experience (e.g., soft–loud, shrill; hereafter we will use '–' to indicate semantically opposite and '/' to indicate semantically similar adjectives). Most other adjectives useful for describing sound, for example, high–low, dark–bright, or rough–smooth, are based on crossmodal associations with other sensory or experiential modalities (Wallmark, 2018; Saitis et al., 2020). In fact, research has demonstrated pronounced crossmodal associations between sound and other modalities such as touch that extend across cultures (Ćwiek et al., 2022; Winter et al., 2022). The study of semantic dimensions of sound has therefore important implications for understanding the relation between perception, language, and meaning.

When it comes to assessing sounds along semantic scales, studies have repeatedly found low dimensional factor spaces with around three dimensions (Zacharakis et al., 2014; Saitis and Weinzierl, 2019). Factors are latent variables resulting from dimensionality reduction techniques such as exploratory factor analysis and principal components analysis (Osgood, 1952). This finding suggests either a limited dimensionality of sound semantics or limited agreement among participants. Little is known about the internal consistency of chatbots such as CGPT. Users quickly grasp that identical prompts can give rise to different responses, but the degree of this variability remains to be quantified.

In this work, we probed CGPT using semantic scales from a recently published study of sound semantics

---

[1]https://openai.com/blog/CGPT/

[2]https://openai.com

using musical instrument tones. Specifically, we had the chatbot provide ratings of sounds along semantic scales, allowing us to directly compare data from human participants with the chatbot. Collecting multiple complete datasets from the chatbot allowed us to compute the consistency of responses. As described in the following, it turned out that the chatbot only agreed with human raters for a selected number of scales. Its internal consistency was remarkably low and in fact comparable to the inter-subject consistency of humans.

2 Methods

2.1 Human Dataset

We considered a recently published dataset by Reymore et al. (2023): 540 online participants rated notes from eight Western orchestral instruments (see Fig. 1) across three registers (low, medium, and high) on 20 semantic scales (5-point). Each scale is described by up to three words: 1) *deep, thick, heavy*; 2) *smooth, singing, sweet*; 3) *projecting, commanding, powerful*; 4) *nasal, buzzy, pinched*; 5) *shrill, harsh, noisy*; 6) *percussive (sharp beginning)*; 7) *pure, clear, clean*; 8) *brassy, metallic*; 9) *raspy, grainy, gravelly*; 10) *ringing, long decay*; 11) *sparkling, brilliant, bright*; 12) *airy, breathy*; 13) *resonant, vibrant*; 14) *hollow*; 15) *woody*; 16) *muted, veiled*; 17) *sustained, even*; 18) *open*; 19) *focused, compact*; 20) *watery, fluid*.

These scales were derived from interviews and rating tasks for imagined typical sounds of instruments (Reymore and Huron, 2020). They are thus well suited to the task: CGPT cannot listen to a musical tone but might have a general understanding of how it sounds. Indeed, after providing its ratings, CGPT added that these "are based on general observations and typical characteristics of a [instrument]'s sound in different registers."

2.2 CGPT Dataset

We designed the following prompt template for CGPT: *Please rate how well the descriptions provided in the following (separated by semi-colons) apply to the sound of a [instrument] for a low-register, mid-register, and high-register note on a scale from 1 to 5 where 1 is "does not describe at all", 3 is "describes moderately well" and 5 is "describes extremely well": 1) deep, thick, heavy; 2) smooth, singing, sweet; ... 20) watery, fluid. Provide your response in matrix form, with columns corresponding to registers (low, mid, high) and rows corresponding to attributes 1-20 (in matched order).* The last bit was necessary to ensure that the model both provided numerical values and in a format suitable for data collection.

We prompted CGPT in early and middle February 2023 using its online public interface[3]. For each triplet of stimuli (one instrument × three registers) we collected 50 ratings from CGPT in an equal number of separate conversations with the chatbot, considering each conversation as a unique "rater." This ruled out CGPT remembering past ratings—the model is not able to access past conversations to inform its responses to the current conversation—while at the same time allowing us to assess the internal consistency of its ratings and compare it with human ratings. Data from four CGPT "raters" were discarded due to many NAs.

2.3 Analysis

To evaluate how CGPT assesses sound semantics, we first computed Pearson correlation coefficients between the model's ratings and those made by humans. Subsequently, we quantified the reliability of the model by computing Pearson correlations between all pairs of responses within each group. In addition, we used exploratory factor analysis to compare latent semantic representations between humans and CGPT.

3 Results

To assess the role of musical training in our sample of human participants, we correlated the average rating profile from musician and nonmusician participants. In the original study of Reymore et al. (2023) musical training was self-reported using the single-question item from the Ollen Musical Sophistication Index (Ollen, 2006): 81% of the participants self-identified as nonmusicians ("non-musician" and "music-loving non-musician") and 19% as musicians ("amateur," "serious amateur," "semiprofessional," and "professional"). Average ratings from these two groups of participants turned out to be highly correlated (Pearson's $r = .97, CI : [.96, .97], p < .001$). For that reason, all 540 human participants were considered jointly in the following.

An important first comparison between human and machine ratings concerns the average rating profiles. Figure 1 shows the average rating profiles, with varying degrees of agreement between human and CGPT ratings. For instance, the scale *deep, thick, heavy* yielded a strong correlation of $r = .86$ [.82, .88] between average human and machine ratings (square brackets correspond to 95% confidence intervals obtained by bootstrapping the CGPT data). With somewhat less agreement, CGPT ratings for the scale *sparkling, brilliant, bright* with $r = .81$ [.73, .87] showed similar patterns across registers but a marked offset towards higher rating points compared to human ratings. Overall, twelve scales yielded positive correlations between human and CGPT ratings with confidence intervals non-overlapping with zero (deep, nasal, shrill, brassy, raspy, ringing, sparkling,
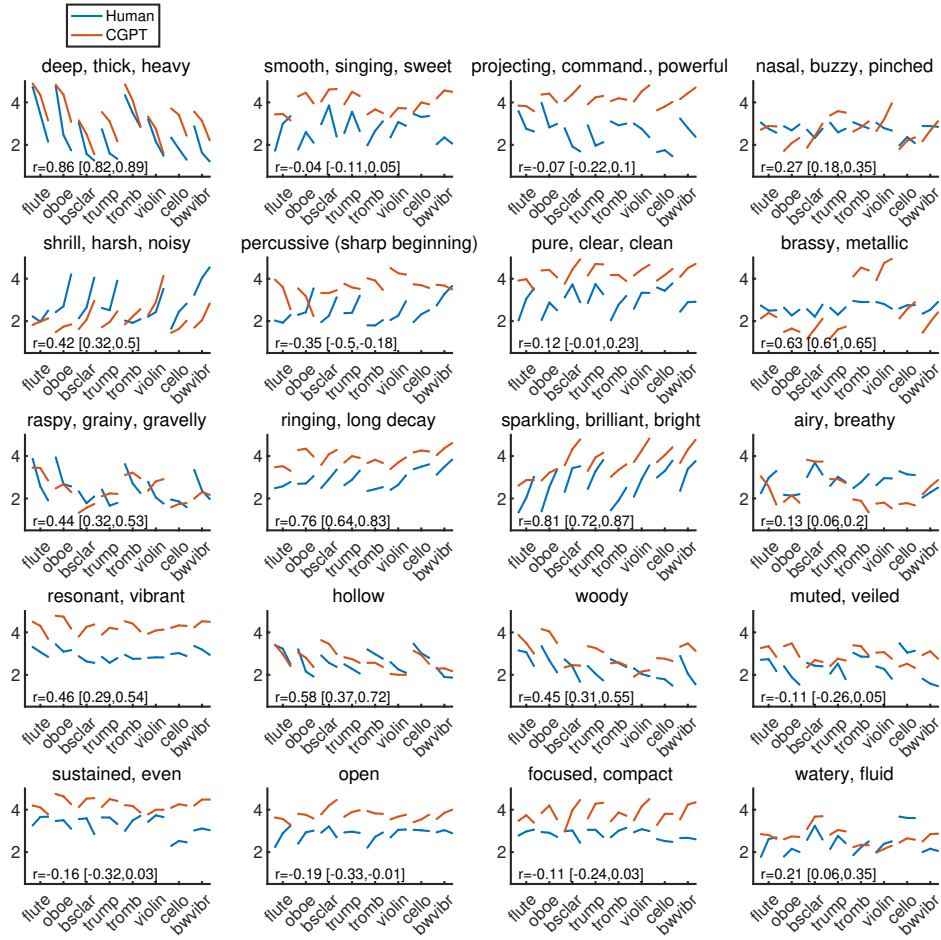
---

[3]https://chat.openai.com/

Figure 1: Profiles of average ratings across all 24 stimuli for human and machine ratings.
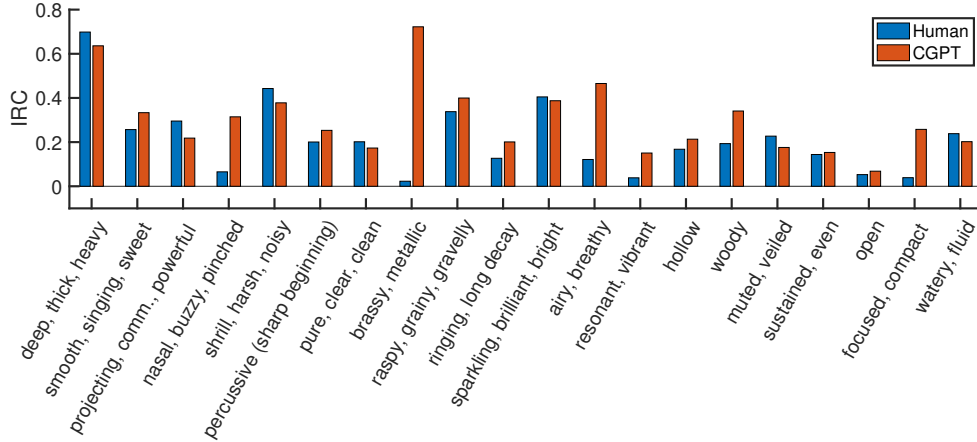
Figure 2: Median inter-rater correlations (IRC) for human and CGPT responses for all 20 semantic scales.

airy, resonant, hollow, woody, and watery).

Next, we considered the internal consistency of human and machine ratings. For every scale, we computed the Pearson correlation between the individual rating profiles of all pairs of participants. The resulting distributions of inter-rater correlations (IRC) correspond to the consistency of responses within the two participant groups. Figure 2 shows the distributions' median values. Note that we could also have shown the IRC for pairs of human and CGPT responses, but these information are redundant (and strongly correlated, r(19) = .79) with the correlations between average rating profiles presented above.

The highest IRC for humans was observed for the *deep, thick, heavy* scale, with similar IRC for CGPT responses. An interesting case is the *brassy, metallic* scale, with an IRC of 0.72 for CGPT responses, but an IRC of close to zero for humans. More generally, calculating the median IRC across scales suggested highest consistency among CGPT responses (0.26) and somewhat lower values for human responses (0.20). Wilcoxon signed rank tests suggested that the medians of the distribution of human-human IRCs (allocated across scales) was only marginally smaller compared to the CGPT-CGPT IRCs ($z = -1.9, p = .057$). That is, ratings from CGPT were similarly consistent compared to human ratings.

To estimate a baseline below which median IRC values could be considered indistinguishable from chance, we conducted a bootstrapping analysis. Median IRC values were computed as before but with randomly shuffled stimulus indices for every participant, using 400 bootstrap iterations. We selected the 99th percentile of the resulting bootstrap distribution of median IRC values as a baseline, amounting to baseline values of 0.0001 and 0.022 for human and CGPT responses, respectively. This implied that indeed all IRC values from all human
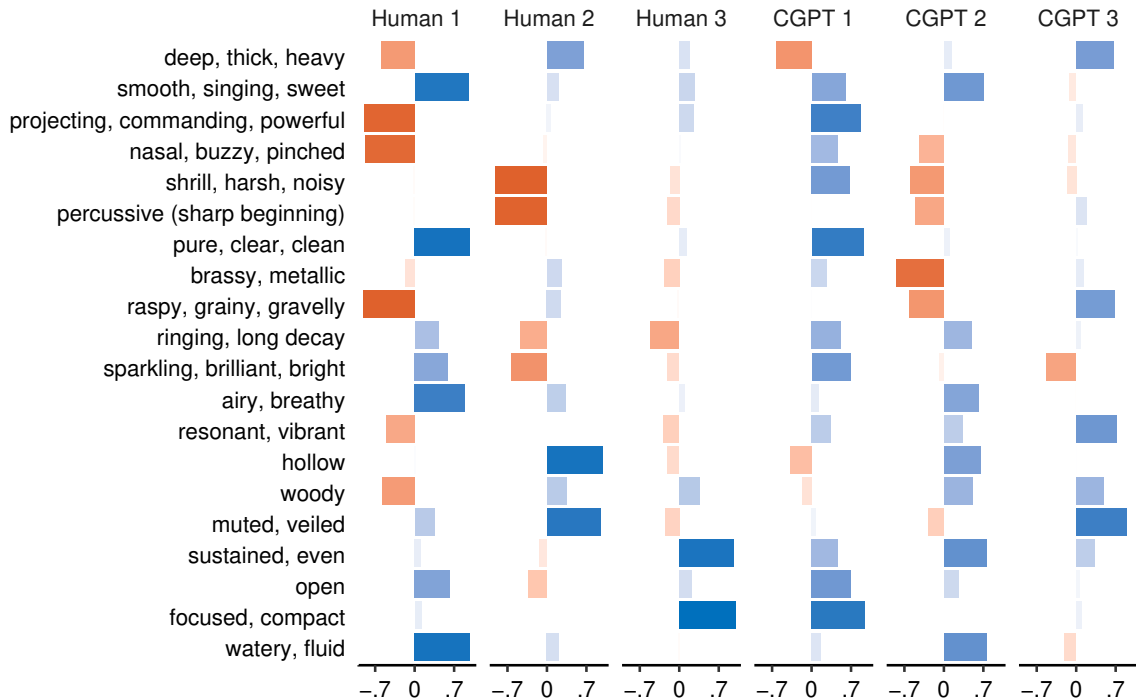
Figure 3: Human and CGPT factor loadings of semantic scales after oblimin rotation.

and CGPT responses were above baseline. Accordingly, even though many of the depicted IRC values are of very small magnitude, they unlikely arose by chance.

We further considered the variance of the average rating profiles and its relation to the IRC. Comparing the average rating profiles and the median IRCs across all 20 scales, we in fact observed very strong correlations between the standard deviations of average rating profiles and the median IRCs within both types of groups (human: $r = .99, p < .001$; CGPT: $r = .98, p < .001$). This implies that in this data set, average rating profiles with little variance were due to inconsistency among participants and not due ratings with little rating variance across stimuli.

Finally, we examined the structure of human and CGPT representations of sound semantics using exploratory factor analysis based on all 20 scales. To select an appropriate number of factors for each dataset, we used a bootstrap method known as Horn's parallel analysis (Horn, 1965). Mardia's test for multivariate normality suggested non-normal distributions in both datasets, therefore factor analysis was performed using principal axis factoring (Fabrigar et al., 1999) with non-orthogonal oblimin rotation. Parallel analysis supported a three-factor solution for both human and machine ratings. The factors cumulatively accounted for 82% and 70% of data variance in the human and CGPT ratings, respectively. Individual factor variance is not available for the rotated solution due to the non-orthogonality of the factors.

The loadings of factors onto semantic scales are shown in Fig. 3. Correlations between individual human and machine factors were moderate to weak, absolute Pearson r values being between .003 (Human 3-CGPT 3) and .38 (Human 3-CGPT 2) with confidence intervals overlapping with zero. The overall correlation between the two semantic spaces was indistinguishable from chance, $r = -.07$ $[-.17, .25]$. That is, even though robust correlations between human and machine ratings were observed for selected scales, there was no correlation between the overall spatial configuration of all 20 scales as derived by factor analysis.

4   Discussion

Here, we took first steps into exploring commonalities and differences in the way in which humans and the chatbot CGPT assess sounds semantics. Specifically, we asked CGPT to perform a recently published rating task of musical instrument sounds along 20 verbally anchored scales (Reymore et al., 2023). Average human and CGPT profiles correlated significantly for twelve out of twenty scales. Specifically, the scales *deep, thick, heavy* and *sparkling, brilliant, bright* yielded strong correlations ($r > .8$). It is worth mentioning that these two scales are related to two of the most salient perceptual dimensions of musical sounds, namely pitch height (*deep–high*) and brightness (*bright–dark*). These dimensions are not completely independent for natural sounds (Siedenburg et al., 2021; Russo and Thompson, 2005), but appear to yield robust agreement between human and CGPT ratings. This is consistent with the findings regarding pitch height by Marjieh et al. (2023). Other less established and potentially less salient semantic scales (e.g., *projecting, …,* or *muted*) accordingly did not yield robust correlations between human and machine responses. Overall, this result suggests a partial agreement between the semantic assessment of sounds by human and CGPT. This was further corroborated by an exploratory factor analysis of the raw human and CGPT semantic ratings data. We found that between the two groups the latent semantic representation captured by the 20 scales shared the same dimensionality but dimensional similarity was as good as chance.

A specific feature of our current approach was to gather multiple responses from CGPT, allowing us to quantify and compare the consistency between responses. We quantified the amount of variability between human raters as well as CGPT responses to the same prompt. This approach showed highest consistency between responses generated by CGPT with higher levels of IRC of the ratings of the chatbot (across different conversations) compared to human ratings (across different participants). Certainly, the reliability of CGPT may be affected by several boundary conditions. For instance, it is specifically designed for chatbot applications, while GPT3 and its successor GPT4 are more general-purpose models and can be used for a wider range of tasks

([Marjieh et al.](), [2023](); [Hansen and Hebart](), [2022]()). Further, the online public CGPT interface, which we used to prompt the chatbot, is more diverse and unpredictable in its output than GPT3 due to a default sampling "temperature" of 1 versus 0.7. This parameter controls how deterministic or "creative" the generated responses are with values between 0 and 1, respectively. (As of early March 2023 it is now possible to specify the CGPT temperature via its API.) It remains an open question to what extent temperature affects the consistency between multiple responses from GPT models when replicating human perceptual tasks (cf. [Miotto et al.](), [2022]()). More generally, it seems to be important for our understanding of CGPT that its ratings show a comparable variance across conversations as responses across human raters.

The pace of development of large language models such as CGPT is extraordinarily high and by the time of writing this manuscript, the model GPT4 has been released. Therefore, the present work can only be considered as a snapshot into how one specific version of CGPT (GPT3, probed in February 2023) construes sound semantics. Yet, researchers might start to track the evolution of machine behavior over time (or development cycles), so that a more complete picture concerning the relation of human and machine semantics can be drawn. Analysing the structure and behavior of deep-learning based algorithms optimized to replicate human perceptual tasks has already become an important approach in perception science ([Eickenberg et al.](), [2017](); [Kell et al.](), [2018](); [Giordano et al.](), [2023]()).

An important difference between the present study and that of [Marjieh et al.]() ([2023]()) concerns the relation between the prompted task and what the model might already know about it. CGPT and GPT3 are both part of a series of models trained on a blend of text and code from before Q4 2021. The human data considered in this study appeared in early 2023, therefore they do not form part of CGPT's "knowledge." Whereas the human loudness and timbre dissimilarity datasets considered by [Marjieh et al.]() ([2023]()) were published in 1978 and 2018, respectively. Is it possible that in their case GPT3 was aware of these studies? And could that have influenced the high correlation they observed for loudness and the moderate correlation for timbre?

Yet another question that could be explored concerns the design of the prompt. Here, we asked the chatbot to provide several ratings as part of one response in order to keep efficiency of data collection on a reasonable level. From a formal viewpoint, however, one could argue that the collection of one data point per prompt may have better fitted with the human experiment. An insurmountable disparity between the way in which data were collected from humans and the chatbot lies in the fact that humans listened to sounds whereas CGPT was simply informed about the instrument name and pitch register. It should be noted that the semantic scales used here were designed on the basis of an auditory imagery task instead of an actual listening task ([Reymore and Huron](), [2020]()). Also, even human judgements of sound similarity have been shown to be highly conflated with

knowledge about instrument categories and the means of sound production (Siedenburg et al., 2016; Saitis and Siedenburg, 2020) and dissimilarity relations between imagined and heard sounds strongly correlate (Halpern et al., 2004).

In summary, the present study addresses the extent to which human sensory judgments can be recovered by machine semantics. It turns out that CGPT agrees with humans with its ratings of a majority of scales including the perceptually most salient ones. In his taxonomy of listening, Schaeffer (2017) suggested four distinct modes of listening: *to listen* (écouter), *to perceive aurally* (ouïr), *to hear* (entendre), *to understand* (comprendre). In the era of artificial intelligence, one might be tempted to suggest yet another mode that captures what CGPT is doing rather successfully already: *to pretend to understand*. The fundamental gap between a listening participant and a non-listening machine remains, but this gap might be closed by future work on machine listening.

Acknolowledgements

References

Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, et al. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, 377(1841):20200390, 2022.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272, 1999.

Bruno L Giordano, Michele Esposito, Giancarlo Valente, and Elia Formisano. Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, pages 1–9, 2023.

Andrea R Halpern, Robert J Zatorre, Marc Bouffard, and Jennifer A Johnson. Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, 42(9):1281–1292, 2004.

Hannes Hansen and Martin N Hebart. Semantic features of object concepts generated with GPT-3. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, pages 779–786, 2022.

John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30:179–185, 1965.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98:1–15, 2018.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. What language reveals about perception: Distilling psychophysical knowledge from large language models. *arXiv preprint arXiv:2302.01308*, 2023.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. Who is GPT-3? An exploration of personality, values and demographics. *arXiv:2209.14338*, 2022.

Joy E Ollen. *A criterion-related validity test of selected indicators of musical sophistication using expert ratings.* PhD thesis, The Ohio State University, 2006.

Charles E Osgood. The nature and measurement of meaning. *Psychological bulletin*, 49(3):197, 1952.

Lindsey Reymore and David Huron. Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia. *Psychomusicology: Music, Mind, and Brain*, 30(3):124–144, 2020.

Lindsey Reymore, Jason Noble, Charalampos Saitis, Caroline Traube, and Zachary Wallmark. Timbre semantic associations vary both between and within instruments: An empirical study incorporating register and pitch height. *Music Perception*, 2023.

F.A. Russo and W.F. Thompson. An interval size illusion: The influence of timbre on the perceived size of melodic intervals. *Attention, Perception, & Psychophysics*, 67(4):559–568, 2005.

Charalampos Saitis and Kai Siedenburg. Brightness perception for musical instrument sounds: Relation to timbre dissimilarity and source-cause categories. *The Journal of the Acoustical Society of America*, 148(4): 2256–2266, https://doi.org/10.1121/10.0002275, 2020.

Charalampos Saitis and Stefan Weinzierl. The semantics of timbre. In Kai Siedenburg, Charalampos Saitis, Stephen McAdams, Arthur N. Popper, and Richard R. Fay, editors, *Timbre: Acoustics, Perception, and Cognition*, pages 119–149. Springer, 2019.

Charalampos Saitis, Stefan Weinzierl, Katharina von Kriegstein, Sølvi Ystad, and Christine Cuskley. Timbre semantics through the lens of crossmodal correspondences: A new way of asking old questions. *Acoustical Science and Technology*, 41(1):365–368, 2020.

Pierre Schaeffer. *Treatise on Musical Objects: An Essay Across Disciplines*, volume 20. Univ of California Press, 2017.

Kai Siedenburg, Kiray Jones-Mollerup, and Stephen McAdams. Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds. *Frontiers in Psychology*, 6(1977): doi: 10.3389/fpsyg.2015.01977, 2016.

Kai Siedenburg, Simon Jacobsen, and Christoph Reuter. Spectral envelope position and shape in orchestral instrument sounds. *The Journal of the Acoustical Society of America*, 149(6):3715–3727, https://doi.org/10.1121/10.0005088, 2021.

Zachary Wallmark. A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, page 0305735618768102, 2018.

Bodo Winter, Márton Sóskuthy, Marcus Perlman, and Mark Dingemanse. Trilled/r/is associated with roughness, linking sound and touch across spoken languages. *Scientific Reports*, 12(1):1035, 2022.

Asterios Zacharakis, Konstantinos Pastiadis, and Joshua D Reiss. An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception: An Interdisciplinary Journal*, 31(4): 339–358, 2014.