# When ChatGPT talks timbre

Charalampos Saitis[1†*] and Kai Siedenburg[2*]
[1] Centre for Digital Music, Queen Mary University of London, London, UK
[2] Department of Medical Physics and Acoustics, University of Oldenburg, Oldenburg, Germany
[†] Corresponding author: c.saitis@qmul.ac.uk [*] Equal contribution

Disciplines: music cognition, musicology, artificial intelligence

Keywords: timbre, semantics, artificial intelligence, natural language processing

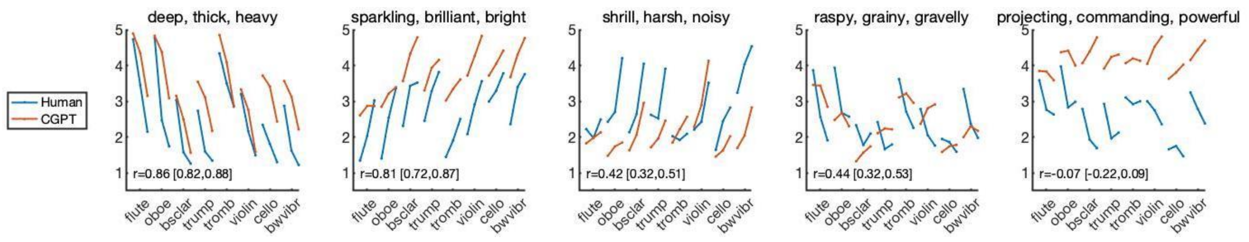
## Introduction

It has recently become obvious that language processing using artificial intelligence is not only here to stay, but also here to change research and society. This insight dawned most saliently with the surfacing of the chatbot ChatGPT (CGPT), based on the large language model (LLM) GPT3 of the company OpenAI. For psychological research, an important question concerns the extent to which LLMs faithfully represent 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 are interested in the extent to what the CGPT chatbot aligns with human judgments of semantic dimensions of sound, the study of which has important implications for understanding the relation between perception, language, and meaning (Saitis & Weinzierl, 2019).

## Methods

We considered the dataset by Reymore et al. (2023): 540 online participants rated notes from eight Western orchestral instruments (flute, oboe, bass clarinet, trumpet, trombone, cello, violin, bowed vibraphone) across three registers (low, medium, and high) on 20 semantic scales (5-point; see Table 1). These were derived from interviews and rating tasks for *imagined typical sounds* of instruments (Reymore & 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, the chatbot added that these "*are based on general observations and typical characteristics of a* [instrument]*'s sound in different registers*." We note that dissimilarity relations between imagined and heard sounds strongly correlate (Halpern et al., 2004), while those between heard sounds can be conflated with knowledge about source-cause categories (Saitis & Siedenburg, 2020; Siedenburg et al., 2016).

Our CGPT prompt template was designed as follows: "*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; ... 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).*" We collected 50 rounds of ratings from CGPT. This allowed 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.



## Results

We first focussed on the five scales that were fitted reasonably well ($R^2 > .2$) by the main LME model in the original publication. Ratings for each scale and each of the 24 stimuli were averaged across raters in each group. Correlation between human and machine ratings (Fig. 1) was high for *deep, thick, heavy* ($r = .86$ [.82, .88]) and *sparkling, brilliant, bright* ($r = .81$ [.73, .87]), with similar patterns across registers but a marked offset towards higher points for machine versus human ratings (square brackets correspond to 95% confidence intervals obtained by bootstrapping the CGPT data). *Shrill, harsh, noisy* and *raspy, grainy, gravelly* both showed positive, albeit weaker correlations ($r > .41$, lower bound of CI > .32). The correlation for *projecting, commanding, powerful* was indistinguishable from chance.

*Inter-Rater Correlation*

Next, we considered the consistency of ratings from human participants and CGPT. For each of the five scales considered here, we computed correlations between all pairs of responses within groups (human-human, CGPT-CGPT) and across groups (human-CGPT). For *deep*, median inter-rater correlations (IRC) were moderate (IRC =.65) and of comparable magnitude in the three types of comparisons. A similar pattern was observed for *sparkling,* albeit with weaker correlations (median IRC = .38). The situation differed for the other three scales, where within-group correlations was generally positive (median IRC > .29), but agreement between ratings from humans and CGPT broke down with IRC close to zero. We found very strong correlations between median IRC and the standard deviations of average rating profiles within both the human ($r = .99$, $p < .001$) and CGPT ($r = .98$, $p < .001$) groups of ratings, implying that scales with little variance of average rating profiles featured little agreement among raters.

*Factor Analysis*

We finally explored human and CGPT timbre semantic spaces using exploratory factor analysis based on all 20 semantic scales. Horn's (1965) parallel analysis supported a three-factor solution for both groups. Factor analysis was performed using principal axis (PA) factoring with non-orthogonal oblimin rotation to account for non-normal multivariate distributions. The factors cumulatively accounted for 82% and 70% of data variance in the human and machine ratings, respectively. Individual factor variance is not available for the rotated solution due to the non-orthogonality of the factors. Human and CGPT factor loadings of the 20 semantic scales are reported in Table 1. Correlations between individual human and machine factors were moderate to weak, absolute Pearson $r$ values being between .003 (Human PA3-CGPT PA3) and .38 (Human PA3-CGPT PA2) with confidence intervals overlapping with zero. The overall correlation between the two semantic spaces was indistinguishable from chance, $r = -.07$ [-.17, .25].

*Table 1: Human and CGPT factor loadings of semantic scales after oblimin rotation.*

| | **Human** | | | **ChatGPT** | | |
|---|---|---|---|---|---|---|
| ***Scales*** | **PA1** | **PA2** | **PA3** | **PA1** | **PA2** | **PA3** |
| Deep, thick, heavy | -.59 | .64 | .19 | -.62 | .13 | .67 |
| Smooth, singing, sweet | **.96** | .20 | .27 | .60 | **.70** | -.12 |
| Projecting, commanding, powerful | **-.89** | .07 | .26 | **.88** | -.01 | .12 |
| Nasal, buzzy, pinched | **-.88** | -.07 | .03 | .46 | -.44 | -.14 |
| Shrill, harsh, noisy | -.02 | **-.92** | -.16 | .69 | -.60 | -.16 |
| Percussive (sharp beginning) | -.02 | **-.90** | -.22 | -.01 | -.51 | .18 |
| Pure, clear, clean | **.98** | -.03 | .13 | **.92** | .09 | .04 |
| Brassy, metallic | -.17 | .25 | -.27 | .27 | **-.84** | .13 |
| Raspy, grainy, gravelly | **-.92** | .25 | -.03 | .01 | -.61 | .68 |
| Ringing, long decay | .42 | -.48 | -.51 | .52 | .48 | .09 |
| Sparkling, brilliant, bright | .59 | -.63 | -.21 | .69 | -.08 | -.53 |
| Airy, breathy | **.90** | .32 | .10 | .13 | .61 | -.01 |
| Resonant, vibrant | -.50 | .00 | -.28 | .34 | .33 | **.71** |
| Hollow | .02 | **.98** | -.21 | -.37 | .65 | .00 |
| Woody | -.58 | .35 | .37 | -.17 | .50 | .50 |
| Muted, veiled | .35 | **.95** | -.25 | .08 | -.28 | **.89** |
| Sustained, even | .12 | -.14 | **.97** | .47 | **.76** | .33 |
| Open | .62 | -.33 | .23 | **.70** | .26 | .06 |
| Focused, compact | .12 | .00 | **1.00** | **.95** | .00 | .09 |
| Watery, fluid | **.98** | .21 | -.02 | .17 | **.76** | -.21 |

*In bold are factor loadings ≥ .70*

## Perspectives

Even though preliminary in nature, this study reveals first insights into the way in which musical timbre is semantically represented by large language models. 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 pitch height (*deep–high*) and brightness (*bright–dark*). These dimensions are not completely independent for natural sounds (Siedenburg et al., 2021; Russo & 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). Unexpectedly, the chatbot showed degrees of internal variability that were comparable in magnitude to that of human ratings. Certainly, the reliability of CGPT may be affected by several boundary conditions (e.g., it is explicitly designed for use in chatbot applications). Yet it seems to be important for our understanding of CGPT that its ratings show a comparable variance across conversations as responses across human raters.

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 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? In contrast, the human data considered in this study appeared in early 2023, therefore they do not form part of CGPT's "knowledge."

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).

## Acknowledgments

KS was supported by a Freigeist Fellowship of the Volkswagen Foundation. The authors thank Lindsey Reymore for making the data of her original study openly available. The authors also thank Elif Ozgür for helping collect some of the ChatGPT data.

## References

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152, 184–194.

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. Nature Neuroscience, 26, 664–672.

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

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

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630-644.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2023). What Language Reveals about Perception: Distilling Psychophysical Knowledge from Large Language Models. *arXiv:2302.01308*.

Reymore, L., & Huron, D. (2020). Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia. *Psychomusicology*, *30*(3), 124–144.

Reymore, L., Noble, J., Saitis, C., Traube, C., & Wallmark, Z. (2023). Timbre Semantic Associations Vary Both Between and Within Instruments: An Empirical Study Incorporating Register and Pitch Height. *Music Perception*, *40*(3), 253–274.

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

Saitis, C., & Siedenburg, K. (2020). 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.

Saitis, C., & Weinzierl, S. (2019). The Semantics of Timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), Timbre: Acoustics, Perception, and Cognition (pp. 119–149). Springer.

Siedenburg, K., Jacobsen. S., & Reuter, C. (2021). Spectral envelope position and shape in orchestral instrument sounds. *The Journal of the Acoustical Society of America, 149*(6), 3715–3727.

Siedenburg, K., Jones-Mollerup, K., & McAdams, S. (2016). Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds. *Frontiers in psychology*, *6*, 1977.