

# GENDER-CODED SOUND: ANALYSING THE GENDERING OF MUSIC IN TOY COMMERCIALS VIA MULTI-TASK LEARNING

Luca Marinelli      György Fazekas      Charalampos Saitis

C4DM, Queen Mary University of London, UK

{l.marinelli, c.saitis, george.fazekas}@qmul.ac.uk


## ABSTRACT

Music can convey ideological stances, and gender is just one of them. Evidence from musicology and psychology research shows that gender-loaded messages can be reliably encoded and decoded via musical sounds. However, much of this evidence comes from examining music in isolation, while studies of the gendering of music within multimodal communicative events are sparse. In this paper, we outline a method to automatically analyse how music in TV advertising aimed at children may be deliberately used to reinforce traditional gender roles. Our dataset of 606 commercials included music-focused mid-level perceptual features, multimodal aesthetic emotions, and content analytical items. Despite its limited size, and because of the extreme gender polarisation inherent in toy advertisements, we obtained noteworthy results by leveraging multi-task transfer learning on our densely annotated dataset. The models were trained to categorise commercials based on their intended target audience, specifically distinguishing between masculine, feminine, and mixed audiences. Additionally, to provide explainability for the classification in gender targets, the models were jointly trained to perform regressions on emotion ratings across six scales, and on mid-level musical perceptual attributes across twelve scales. Standing in the context of MIR, computational social studies and critical analysis, this study may benefit not only music scholars but also advertisers, policymakers, and broadcasters.

## 1. INTRODUCTION

The purpose of this study is to analyse gender-coding in a context where music is secondary to other modalities and serves a clear purpose, such as in advertisement. Our aim is to investigate how music may be employed to reinforce traditional gender roles in toy commercials, and we propose an automatic method for analyzing this phenomenon.<sup>1</sup> Our overarching research objective is to pro-

<sup>1</sup> [https://github.com/marinelliluca/gender\\_coded\\_sound\\_ismir2023](https://github.com/marinelliluca/gender_coded_sound_ismir2023)

 © L. Marinelli, C. Saitis, and G. Fazekas. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** L. Marinelli, C. Saitis, and G. Fazekas, "Gender-coded sound: Analysing the Gendering of Music in Toy Commercials via Multi-task Learning", in *Proc. of the 24th Int. Society for Music Information Retrieval Conf.*, Milan, Italy, 2023.

vide a basis for a theory of message production. Specifically, a theory of the effects that message producers, their decision-making, or their unconscious gender biases have on the selection and composition of sound and music in toy adverts. For this goal, we propose an integrative approach combining content analytical (CA) variables, music perceptual ratings, and multimodal affective ratings to annotate toy commercials, and using multi-task learning (Fig. 1) to analyse the gendering of their soundtracks.

### 1.1 Gendered music styles as cognitive schemas

Empirical studies have demonstrated that gender and sex impact the perception and processing of music [1–3]. However, the idea that sex determines fixed differences in brain structure has been questioned due to potential misinterpretations, overestimations, and publication bias [4]. Gender schemas, instead, are *learned* cognitive networks of associations that guide an individual's behavior by assimilating or rejecting gender-appropriate ideas and activities [5,6]. Schemas guide an individual's perception, information processing, and memory retention, as they prevent information overload by organising one's perceptual experience into a coherent and intelligible whole [6,7].

Popular music genres have been themselves theorised as cognitive schemas containing extramusical concepts that can be primed when a subject is exposed to the genre's music [8,9]. Such schemas are formed through repeated exposure to the multimodal discourse encompassing music, which is to some extent globalised, but that also varies from culture to culture as a result of glocalisation.<sup>2</sup> processes [9] Schema theory has also been used in literary reading and analysis to explain organised "bundles of information and features" [10, p. 106-7] such as literary genres (e.g., science fiction, fantasy, horror).

While gendered language patterns in text and lyrics [11] may be relatively more straightforward to interpret, gender-coding in sound and music ensues from the historical sedimentation, in musical practice, of multimodal associations between gendered meanings in language, visual images, and musical structures [12]. For example, instruments have been consistently associated with masculinity or femininity, even when their sound is presented in isolation and not visually linked to the actual object [13,14]. Sergeant and Himonides [15,16] investigated whether in Western art music individual sounds or their organization

<sup>2</sup> A lexical blend of globalization and localism.

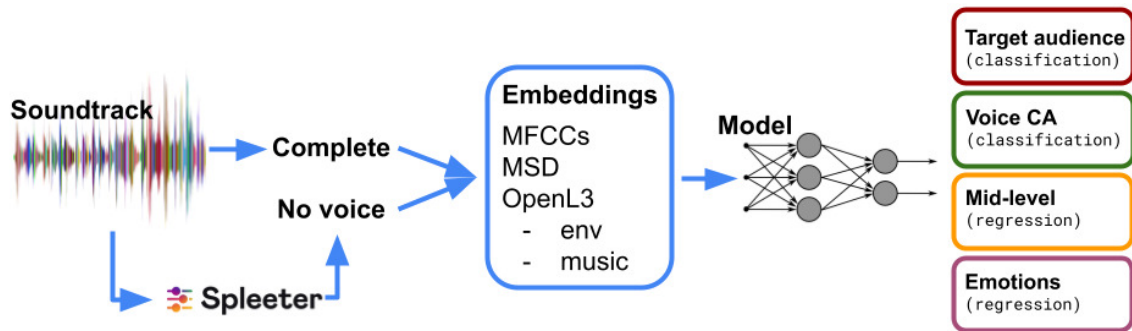


Figure 1. Brief overview of the experimental pipeline.

within a composition could infer the sex or gender of the performer or composer. Even though they found no correlation between the gender of the composers or performers and the gendering of music, raters agreed on the gendering of music, which was associated with features such as tempo, minor/major key, and tonal weight or density. Tagg [17] studied the reception of gendered meanings in TV theme tunes and also found high agreement among participants. Several musical dimensions, such as average tempo, rhythmic and dynamic regularity, and presence of active bass lines, may contribute to conveying gendered meanings. In a subsequent investigation, Tagg and Clarida [18] found that musical pieces linked to female characters were more prone to be classified as quiet and calm. Wang and Horvát [19] computationally extracted twelve descriptors of musical parameters and perceptual features for over 200k songs by more than 8k globally distributed artists across a multitude of popular music genres. They found statistically significant differences for eleven out of twelve musical parameters with regard to the gender of the composers, suggesting the existence of measurable, supra-genre, gendered music styles in the global music industry.

Some of these studies appear to contradict each other,<sup>3</sup> while at the same time sharing the same fallacy, in that feminine and masculine patterns in the performance and composition of music should be considered on a par with distinct gendered styles in spoken language, such as Lakoff’s ‘women’s talk’ [20]. As such, these differences should not be understood in terms of a causal relationship between the gender of the artists and gendered musical patterns. Individuals may have a tendency to use forms of expression that they deem appropriate with regard to their identity, but given the performative nature of gender we cannot possibly generalise this behaviour (i.e., even strong correlation is not causation), as this would end up reinforcing gender stereotypes and their power relations.

Gender schemas therefore mediate our perception of music, and this relationship appears to be bidirectional. At the same time, music-primed schemas can alter our perception of other people’s ethnicity, rural/urban background, age, expertise [8], and even gender [21, 22]. We thus posit that not only gender roles and stereotypes can

be understood in terms of schemas, but also that masculine and feminine music styles can be viewed as *music-primed gender schemas*, which to some extent overlap with the former. We also presume that different music-primed schemas might exist for other intersectional factors, such as class.

## 1.2 Gendered toy marketing

Gender polarisation in TV advertising aimed at children has been consistently found in a large body of studies spanning over 40 years [23–25]. Differences in commercials targeted at girls, boys, and mixed audience have been found in terms of: sound (voices, background music and sound effects), language, transitions and camera work, setting, interactions and activities, and colours.

Specifically in terms of sound and music, Welch et al. [23] noted that in general the sex of the voice-over matched the target audience of the commercials, but that male narrating voices also occurred more often in mixed audience commercials, and subsequent research confirmed the same trend [25]. They also found that commercials targeted at boys had more noise, louder music, and more sound effects. Another study [24] conversely found that music used in girls’ advertisements is generally softer and more likely to have a sung narration style. Whereas, Johnson and Young [26] identified what they called “gender exaggeration:” male voice-overs tend to be exceedingly deep, growl-like or aggressive, whereas female voice-overs are often very high-pitched and singsong.

By interpreting music as an inherently multimodal discourse, a critical analysis of gender markers in children’s TV adverts can help to investigate the relation between music and hegemonic discourses on gender; and to promote further research towards a commercial and contemporary musical semiotics of gender. Analysing music in gendered advertising aimed at children allows a privileged glance into the birthplace of music-primed gender schemas.

## 1.3 Automatic discourse processing

Discourse analysis is an umbrella term that refers to approaches developed across diverse academic disciplines. This includes disciplines that first developed models for understanding discourse, such as linguistics, social semiotics and conversation analysis. But it also refers to other

<sup>3</sup> [19] found significant correlations between the gender of the composers and characteristics of their music, while [16] did not.

All <i>N</i> = 606		Feminine <i>N</i> = 163	Masculine <i>N</i> = 149	Mixed <i>N</i> = 200
	<i>Type</i>	$\chi^2(6, N = 512) = 89.02, p = .000$		
5.6%	Sung	9.8%	None	2.5%
18.8%	Spoken and sung	36.8%	6.0%	17.0%
67.7%	Spoken	52.8%	81.2%	75.5%
7.9%	No voices	0.6%	12.8%	5.0%
	<i>Age</i>	$\chi^2(6, N = 512) = 39.51, p = .000$		
79.5%	Adults	76.7%	79.2%	83.0%
5.9%	Children and adults	8.0%	6.0%	6.5%
6.6%	Children	14.7%	2.0%	5.5%
7.9%	No voices	0.6%	12.8%	5.0%
	<i>Gender</i>	$\chi^2(6, N = 512) = 332.1, p = .000$		
39.8%	Feminine	95.7%	2.0%	29.5%
46.9%	Masculine	1.8%	83.9%	54.5%
5.4%	Feminine and masculine	1.8%	1.2%	11.0%
7.9%	No voices	0.6%	12.8%	5.0%
	<i>Gender exaggeration</i>	$\chi^2(6, N = 512) = 243.6, p = .000$		
16.5%	Exagg. feminine	44.8%	None	8.0%
15.5%	Exagg. masculine	None	40.3%	7.0%
60.1%	All normal sounding	54.6%	47.0%	80.0%
7.9%	No voices	0.6%	12.8%	5.0%

**Table 1.** Contingency tables of voice-related content analytical variables with  $\chi^2$  tests of independence. The column "All" includes commercials without actors or presenter (94).

approaches that apply and extend these models of understanding to their particular academic field, such as cognitive psychology, literary criticism and *artificial intelligence* [27]. Research on discourse processing, an endeavour of natural language processing (NLP), is already at a stage where machine learning approaches are able, for example, to automatically detect social attitudes and political stances in online news or social media [28, 29].

Beyond textual discourse and NLP, denotative meanings in images and videos can be easily captured by machine learning techniques [30, 31]. However, works that try to address connotative meanings or the rhetoric of multimedia content are still in their infancy and such approaches are often not even framed as pertaining to discourse or semiotic analysis. Dinkov et al. [32] predicted the political ideological bias (left, centre, right) of media outlets using text, metadata, and audio (via speech processing techniques) from YouTube channels, but not visual content. Ye et al. [33] predicted the messages that image and video advertisements convey by explicitly modeling symbolic associations (e.g., gun for “danger”) and combining cues from multiple modalities, including the loudness in video soundtracks. Notably, none of these studies leveraged approaches and tools from music information retrieval.

### 1.4 Multi-task learning in MIR

In multi-task learning we train a single model to perform multiple related tasks simultaneously, leveraging shared information among tasks, which results in several benefits. Böck et al. [34] simultaneously modelled tempo estima-

tion and beat tracking of musical audio, showing state-of-the-art performance for both tasks. Wu et al. [35] combined multi-task and self-supervised learning, resulting in improved performance. Chowdhury et al. [36] proposed a VGG-style deep neural network to predict emotional characteristics of music based on mid-level perceptual features (e.g., melodiousness and tonal stability) and found that the loss in performance was negligible when compared to predicting emotions directly. Further improvements were obtained by training jointly on the mid-level and emotion annotations, with the small loss in performance justified by the gain in explainability of the predictions. Our study expands upon this foundation by incorporating emotions and perceptual features, while also adding more granular structure to facilitate a comprehensive understanding of the gendering of music in multimodal contexts.

## 2. DATASET

Our hierarchical data collection framework comprised CA variables at the lower level, music-focused ratings from experts at the middle level, and multimodal affective ratings at the highest level of subjectivity. Mid-level perceptual features, which describe relevant and instantly identifiable musical characteristics, exhibit high consistency across listeners and can be predicted from the acoustic signal. These features also correlate with music’s affective dimensions [37]. The emotion ratings were collected from adults rather than children because adults are better equipped to capture the commercials’ intended emotional impact. Fur-

thermore, research indicates that children exhibit adult-like emotion recognition capabilities by age 11 [38].

## 2.1 Sampling method

In March 2022, we collected a sample of 5614 videos from the official YouTube channel of Smyths Toys Superstores, a major UK toy retailer. To ensure comparability with previous studies [39,40], we selected only high-quality videos intended for television and excluded those without audio, formatted for mobile phones, or with substantial on-screen text. Additionally, we excluded advertisements featuring toddlers and pre-schoolers as these are actually targeted at parents. To minimise duplicates, we removed videos with the same title from our sample.

Given that we are interested in understanding the gendering of sound and music in the toy industry at large, we needed to enforce some balance across gender targets. We thus performed a *preliminary* classification of 1778 commercials based on their intended target audience (feminine, masculine or mixed audience) using simple heuristics regarding the gender of the majority of presenters featuring in the commercial, the colour coding of the video and ultimately the category of the product. This resulted in 780 'feminine', 509 'masculine', and 489 'mixed audience' commercials. A final sample of 606 commercials, spanning over 10 years from 2012 to 2022, was obtained by randomly sampling from each category 202 videos.

## 2.2 Content analysis (manual annotation)

The *gender orientation* (also *target audience*) of the commercials was determined by the gender of the actors/presenters. Following [26], in order to account for tokenism, whenever a presenter of the other gender was included in the background or for just a few seconds, these were considered token gender representations and not explicit market orientations. All fictional characters, even when realistic (e.g. from a video game), were not considered as actors/presenters and the corresponding commercials were coded as having no actors. Whenever commercials featured exclusively character 'dismemberment' (e.g., showing only hands without a face or head) [41] these were also coded as having no actors.

Four distinct items describing the sound of the voices in the commercial were collected using a coding schema based on Verna's research [42]. But unlike the original work, we coded for all voices in the commercial, both diegetic and non-diegetic. The reason for this choice is that there is no way to reliably distinguish between diegetic and non-diegetic sounds purely based on the audio signal. Commercials were coded in terms of *type of voices* ("Spoken", "Sung", "Both spoken and sung", "No voices"), then in terms of *voices age* ("Adults" which included young adults, "Children", "Children and adults", "No voices"), *gender exaggeration* of the voices ("All normal sounding", "Exaggeratedly masc.", "Exaggeratedly fem.", "No voices"), and finally in terms of *voice gender* ("Feminine", "Masculine", "Feminine and masculine", "No voices"). In order to determine the reliability of each variable, 15% of

the commercials was double-coded by two coders independently. For all variables we obtained Krippendorff's alpha levels above .80 (with 'gender orientation' and 'gender of the voices' exceeding .90), and therefore met the standards of reliability required for this type of analysis [43]. Out of 606 commercials analyzed, 163 were targeted at a feminine audience, 149 at a masculine audience, 200 at a mixed audience and 94 featured no actors or presenters. Contingency tables of the voice variables are shown in Table 1.

## 2.3 Music-focused and emotion ratings

Participants in our study were paid between £7 and £8 per hour (depending on their completion time) on Prolific.co. In order to minimise the effects of careless responding, a low-effort metric was computed by summing the length of all long strings for each participant, and those that scored above two standard deviations from the average value were screened out during data collection, as it was performed in batches of 50 participants. For 600 of the videos, we collected between five and six ratings on each music and emotion scale. At an initial stage, the remaining 6 videos were used as controls (i.e., were rated by all participants), but we do not leverage them as such in the current study.

Musically trained participants (at least three years of experience with an instrument) rated the soundtracks of the commercials on 15 music-focused bipolar scales [44, 45]: Electric/Acoustic, Distorted/Clear, Loud/Soft, Many/Few instruments, Heavy/Light, High/Low pitch, Punchy/Smooth, Wide/Narrow pitch variation, Harmonious/Disharmonious, Clear melody/No melody, Complex/Simple rhythm, Repetitive/Non-repetitive, Dense/Sparse, Fast/Slow tempo, and Strong/Weak beat. We collected a total of 4560 ratings from 152 participants from the UK (75 M, 77 F, aged  $40 \pm 14$ ). Given that our focus is on music, but soundtracks consist of speech, music and sound effects, our question was formulated as follows: "The following are a series of perceptual attributes of music. You are asked to evaluate the *music in the background* in terms of the adjectives on each side of the scale."

To annotate the perceived affect of videos, we drew from the aesthetic emotions scale [46, AESTHEMOS], which was devised from an extensive review of emotion measures from different domains such as music, literature, film, painting, advertisements, design, and architecture, and is thus ideal, in its flexibility, for our use with multimodal stimuli. Given that our focus is on music and sound, in a preliminary study we limited our choice to a subset of 10 AESTHEMOS items that intersect with the 13 music emotions listed by Cowen et al. [47]. Of these, we kept only seven scales which showed significant discriminant capabilities: Happy or Delightful, Amusing or Funny, Beauty or Liking, Calm or Relaxing, Energising or Invigorating, Angry or Aggressive, and Triumphant or Awe-inspiring. We used a single unipolar item for each subscale, instead of two. We collected a total of 4530 ratings from 151 participants from the UK (76 M, 75 F, aged  $39 \pm 13$ ). Given that our aim is to analyse the intended emotional profile, our question was formulated as follows:

Embeddings	Voice	Target F1	Secon. F1	Avg. $R^2$ emo	Avg. $r$ emo	Avg. $R^2$ mid	Avg. $r$ mid
mfcc	no	.79 ± .08	.66 ± .07	.02 ± .17	.36 ± .11	.14 ± .16	.48 ± .09
mfcc	yes	.78 ± .10	.65 ± .07	.06 ± .16	.38 ± .11	.13 ± .15	.43 ± .10
msd	no	.87 ± .05	.66 ± .06	.25 ± .11	.54 ± .08	.35 ± .14	.62 ± .09
msd	yes	.95 ± .04	.79 ± .05	.26 ± .15	.56 ± .09	.30 ± .12	.58 ± .09
openl3_env	no	.91 ± .05	.72 ± .06	.34 ± .10	.61 ± .08	.41 ± .10	.66 ± .07
openl3_env	yes	.95 ± .04	.77 ± .05	.34 ± .13	.62 ± .08	.35 ± .12	.62 ± .08
openl3_music	no	.87 ± .09	.71 ± .06	.31 ± .16	.56 ± .19	.39 ± .16	.64 ± .15
openl3_music	yes	.91 ± .11	.76 ± .10	.29 ± .17	.56 ± .19	.31 ± .14	.59 ± .13

**Table 2.** Mean and standard deviation from 5x repeated 5-fold cross-validation. 'Target' refers to the gender orientation of ads (binary); secondary tasks involve voice-related content analytical variables. 'No' represents models trained on voice-separated accompaniments, while 'Yes' indicates models trained on entire soundtracks.

"Toys commercials are targeted at an audience mainly consisting of children and aim at evoking the following emotions. Pay attention to both sound and images and rate each *intended* emotion accordingly."

### 2.4 Between-targets ANOVA

We first performed between-targets (i.e., gender targets of the commercials) one-way analyses of variance for each of the music-focused and emotion scales. When ANOVA assumptions were violated, we performed a Kruskal-Wallis H-test instead. Highly significant polarisation ( $p < .001$ ) emerged for twelve of the mid-level music perceptual scales, with stark contrasts observed between feminine and masculine-targeted commercials, and commercials targeted at mixed audiences generally registering in-between values. Masculine adverts were more Electric than Acoustic, more distorted, disharmonious and with a less clear melodic contour than feminine ones. They also were more dense in terms of instrumentation, more Punchy, with stronger beats, and therefore were generally louder and heavier. Also in terms of rhythmic complexity, they were more complex than feminine-targeted commercials. Thus a clear picture emerges, as the soundtracks in boys' adverts are significantly more *abrasive* than those in girls' ads.

Similarly, stark contrasts ( $p < .001$ ) were observed between feminine and masculine-targeted commercials for all affective scales, with commercials targeted at mixed audiences often registering in-between values. Commercials targeted at boys were the least "Happy or delightful", the least "Amusing or funny", "Calm or relaxing", and registered the lowest values on the scale "Beauty or liking". They instead were the most "Energising or invigorating", "Angry or aggressive", and "Triumphant or awe-inspiring". Apart from the scale "Amusing or funny", which scored the highest values within mixed audiences commercials, adverts targeted at girls displayed an opposite behaviour to those for boys. For example, they were the most "Calm or relaxing" and the least "Angry or aggressive". As previously highlighted with the music-focused scales, masculine-targeted commercials appear again to be significantly more *abrasive* than the feminine ones.

We report a more in-depth analysis in an upcoming

publication. In this paper, we exclude "Amusing or funny" from further analyses due to poor correlation with the mid-level features. We also exclude the three non-significant mid-level scales: Wide/Narrow pitch variation, Repetitive/Non-repetitive, and Fast tempo/Slow tempo.

### 3. MACHINE LEARNING PIPELINE

Our machine learning framework is a multi-task learning model implemented in PyTorch (Fig. 1). It was trained to simultaneously learn mid-level features regression, emotion regression, and all the CA variables (classes). These tasks share an initial hidden layer with 128 units and then branch out into separate sub-tasks. Each sub-task has its own hidden layer with 128 units and an output layer with dimensions corresponding to the specific task.

To avoid the jingle of the retailer in the last 5 seconds of most soundtracks, we trimmed them accordingly. Then with Spleeter [48] we separated voices and accompaniments. Features were extracted in non-overlapping chunks across the trimmed soundtrack and then averaged across the chunks. We computed 20-band MFCCs using *librosa* [49], along with their delta and delta-deltas, yielding 60-dimensional embeddings. A reimplementation of a state-of-the-art model trained on the million song dataset (MSD) [50] provided 256-dimensional embeddings. OpenL3 features were computed using the provided *conda* package [51], generating 512-dimensional embeddings for both environmental and music models.

The proposed model employs an equally weighted, combined loss function, incorporating the mean squared error for the mid-level features and emotion regression tasks, and cross-entropy loss for the classification tasks. The model was trained jointly on all tasks. We also used a model checkpoint and early stopping with a patience of 30 epochs (maximum of 200). Repeated 5-fold cross-validation was performed (10% test, 10% validation, for 5 repetitions, i.e. 25 "folds", as the random seed was not set) and utilised the AdamW optimizer instead of Adam for regularization. Further optimising the network to surpass the already remarkable results, as well as conducting ablation studies to evaluate the various components and design choices, is beyond the scope of our investigation.

Embeddings	Voice	Target F1	Secon. F1	Avg. $R^2$ emo	Avg. $r$ emo	Avg. $R^2$ mid	Avg. $r$ mid
mfcc	no	.52 ± .05	.67 ± .06	.04 ± .16	.37 ± .11	.14 ± .14	.48 ± .09
mfcc	yes	.48 ± .04	.67 ± .05	.05 ± .15	.38 ± .11	.15 ± .14	.46 ± .09
msd	no	.62 ± .05	.67 ± .06	.23 ± .15	.54 ± .10	.36 ± .12	.64 ± .07
msd	yes	.67 ± .06	.80 ± .05	.29 ± .12	.57 ± .08	.33 ± .10	.60 ± .07
openl3_env	no	.59 ± .06	.72 ± .06	.30 ± .12	.59 ± .08	.42 ± .10	.66 ± .07
openl3_env	yes	.66 ± .07	.77 ± .06	.34 ± .11	.61 ± .07	.35 ± .10	.62 ± .07
openl3_music	no	.64 ± .07	.73 ± .06	.32 ± .12	.60 ± .08	.43 ± .11	.68 ± .07
openl3_music	yes	.67 ± .04	.78 ± .05	.35 ± .12	.61 ± .08	.37 ± .10	.63 ± .07

**Table 3.** Same as Table 2, but results refer to ternary ‘Target’ classification.

#### 4. RESULTS

Tables 2 and 3 reveal once again stark differences between the soundtracks of commercials designed for feminine and masculine audiences (the value “no” corresponds to models trained on the voice-separated accompaniments). In fact, the binary classification task on the soundtracks including voice achieves an impressively high Target F1 score of  $.95 \pm .04$  using the MSD and OpenL3 env embeddings. It is also worth noting that even without voice, the soundtracks still contain enough information to classify the commercials with a high degree of accuracy, with the OpenL3 env embedding achieving a Target F1 score of  $.91 \pm .11$ . In a way, the dataset is so gendered that it can be considered a toy dataset in all senses.

Upon closer examination of the  $R^2$  and  $r$  emotions metrics, we observe that they are relatively low across all experiments compared to mid-level metrics. This contrasts with previous research [36] where mid-level correlations were lower than those of emotions, as in our case the  $R^2$  mid-level and  $r$  mid-level metrics are generally higher, with the OpenL3 embeddings performing the best.

When comparing Tables 2 and 3, the high performance of the MSD and both OpenL3 embeddings on the binary task, suggests that there are no significant differences in the soundtracks of mixed-audience commercials compared to those targeted at feminine or masculine audiences. This confirms the results from the analysis of variance and highlights the ability of these embeddings to perform similarly across different target audiences. Overall, we found that the OpenL3 embeddings performed better than others across all tasks, indicating superior generalizability, as already shown in previous research especially in the context of limited training examples [52]. However, the relatively low  $R^2$  and  $r$  for emotions suggest that there is still room for improvement, possibly through multimodal fusion.

It is noteworthy that human voice plays a critical role in conveying higher-level connotations, as performance on the classification tasks *and* especially the emotion regressions generally improves when voices are present. Additionally, improvement in mid-level regressions on the accompaniments of the soundtracks (no voice) indicates that participants in the data collection were able to focus on the background of the soundtracks, as they were asked to.

Although the MFCCs are the worst performing, their

discriminative power on the target task and the decent performance on the mid-level features regression highlight the underlying “simplicity” of the task, in terms of the strong collinearity due to the degree of gender-polarization inherent in the dataset.

#### 5. CONCLUSION

By examining the performance of different musical embeddings in classifying commercials targeted at different audiences, and by providing explainable inference of the target of the commercials, in terms of affective and of music perceptual features, this study sheds light on the role of music in gendered marketing strategies. Such approach has significant implications for advertisers, policymakers, and broadcasters, who recently faced a public backlash against the gendered marketing of toys and other products.<sup>4</sup> Furthermore, the study highlights the importance of considering the role of music when regulating marketing strategies and developing more inclusive and diverse advertising campaigns. Our results suggest that gendered music styles in toy commercials emerge as a result of deliberate marketing strategies, as such styles reflect gender stereotypes that are “ludicrously old-fashioned and offensively out of touch” [53] and still prevalent in the industry.

By bringing together music analysis, machine learning, and critical analysis, this study illustrates the potential of interdisciplinary approaches, contributing to the emerging field of computational social studies. It highlights the importance of considering the role of music, among other modalities, in shaping societal norms and values and the need for greater awareness and accountability in the use of such affordances in marketing and other industries.

Future research can build on these findings by further investigating the relationship between gendered music and advertising strategies in different industries and contexts, exploring the impact of gendered music on consumer behavior and societal perceptions of gender, and developing new methodologies for creating more inclusive and diverse marketing campaigns. The results also emphasise the potential for the development of multimodal approaches to enhance the models’ performance on these tasks.

<sup>4</sup> <https://www.bbc.co.uk/news/world-us-canada-46613032>

## 6. ACKNOWLEDGMENTS

This work was supported by UK Research and Innovation [grant number EP/S022694/1], and was partially conducted during the first author's Enrichment Scheme placement at the Alan Turing Institute. The authors would like to express their gratitude to Professor Petra Lucht for her invaluable guidance at early stages of this study.

## 7. REFERENCES

- [1] J. J. Kellaris and S. P. Mantel, "The influence of mood and gender on consumers' time perceptions," *ACR North American Advances*, 1994.
- [2] G. T. Toney and J. B. Weaver, "Effects of gender and gender role self-perceptions on affective reactions to rock music videos," *Sex Roles*, 1994.
- [3] J. Meyers-Levy and R. Zhu, "Gender differences in the meanings consumers infer from music and other aesthetic stimuli," *Journal of Consumer Psychology*, 2010.
- [4] G. Rippon, "Do women and men have different brains?" *New Scientist*, 2019.
- [5] S. L. Bem, "Gender schema theory: A cognitive account of sex typing," *Psychological review*, 1981.
- [6] E. Leung, "Gender schemas," in *Encyclopedia of Personality and Individual Differences*. Springer, 2020.
- [7] M. G. Boltz, "Musical soundtracks as a schematic influence on the cognitive processing of filmed events," *Music Perception*, 2001.
- [8] M. Shevy, "Music genre as cognitive schema: Extramusical associations with country and hip-hop music," *Psychology of music*, 2008.
- [9] S. Kristen and M. Shevy, "A comparison of German and American listeners' extra-musical associations with popular music genres," *Psychology of Music*, 2013.
- [10] P. Stockwell, *Cognitive poetics: An introduction*, 2nd ed. Routledge, 2019.
- [11] L. Betti, C. Abrate, and A. Kaltenbrunner, "Large scale analysis of gender bias and sexism in song lyrics," *arXiv preprint arXiv:2208.02052*, 2022.
- [12] N. Dibben, "Gender identity and music," in *Musical identities*. New York: Oxford University Press, 2002.
- [13] C. A. Elliot and M. Yoder-White, "Masculine/feminine associations for instrumental timbres among children seven, eight, and nine years of age," *Contributions to Music Education*, 1997.
- [14] L. M. Stronsick, S. E. Tuft, S. Incera, and C. T. McLennan, "Masculine harps and feminine horns: Timbre and pitch level influence gender ratings of musical instruments," *Psychology of Music*, 2018.
- [15] D. C. Sergeant and E. Himonides, "Gender and the performance of music," *Frontiers in psychology*, 2014.
- [16] ———, "Gender and music composition: A study of music, and the gendering of meanings," *Frontiers in psychology*, 2016.
- [17] P. Tagg, "An anthropology of stereotypes in tv music?" *Swedish Musicological Journal*, vol. 71, 1989.
- [18] P. Tagg and B. Clarida, "Title tune gender and ideology," in *Ten little title tunes: towards a musicology of the mass media*. Huddersfield: The Mass Media Music Scholars' Press, NY., 2003.
- [19] Y. Wang and E.-Á. Horvát, "Gender differences in the global music industry: Evidence from musicbrainz and the echo nest," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2019.
- [20] R. Lakoff, "Language and woman's place," *Language in society*, 1973.
- [21] H.-B. Brosius and H. M. Kepplinger, "Der einfluß von musik auf die wahrnehmung und interpretation einer symbolisierten filmhandlung," *Rundfunk und Fernsehen*, 1991.
- [22] A.-K. Herget, "On music's potential to convey meaning in film: A systematic review of empirical evidence," *Psychology of Music*, 2021.
- [23] R. L. Welch *et al.*, "Subtle sex-role cues in children's commercials," *Journal of Communication*, 1979.
- [24] J. Lewin-Jones and B. Mitra, "Gender roles in television commercials and primary school children in the uk," *Journal of children and media*, 2009.
- [25] B. Mitra and J. Lewin-Jones, "Colin won't drink out of a pink cup," in *The handbook of gender, sex, and media*. Wiley Online Library, 2012.
- [26] F. Johnson and K. Young, "Gendered voices in children's television advertising," *Critical Studies in Media Communication*, 2002.
- [27] D. Schiffrin, D. Tannen, and H. E. Hamilton, "Introduction to the first edition," *The handbook of discourse analysis*, 2015.
- [28] Y. Feng, H. Chen, and L. He, "Consumer responses to femvertising: a data-mining case of dove's "campaign for real beauty" on youtube," *Journal of Advertising*, 2019.
- [29] G. Wiedemann, "Text mining for discourse analysis: An exemplary study of the debate on minimum wages in Germany," *Quantifying approaches to discourse for social scientists*, 2019.
- [30] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

- [31] S. Islam, A. Dash, A. Seum, A. H. Raj, T. Hossain, and F. M. Shah, "Exploring video captioning techniques: A comprehensive survey on deep learning methods," *SN Computer Science*, 2021.
- [32] Y. Dinkov, A. Ali, I. Koychev, and P. Nakov, "Predicting the leading political ideology of youtube channels using acoustic, textual, and metadata information," *arXiv preprint arXiv:1910.08948*, 2019.
- [33] K. Ye, N. H. Nazari, J. Hahn, Z. Hussain, M. Zhang, and A. Kovashka, "Interpreting the rhetoric of visual advertisements," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [34] S. Böck, M. E. P. Davies, and P. Knees, "Multi-task learning of tempo and beat: Learning one to improve the other," in *International Society for Music Information Retrieval Conference*, 2019.
- [35] H.-H. Wu, C.-C. Kao, Q. Tang, M. Sun, B. McFee, J. P. Bello, and C. Wang, "Multi-task self-supervised pre-training for music classification," in *ICASSP 2021-2021*. IEEE, 2021.
- [36] S. Chowdhury, A. Vall, V. Haunschmid, and G. Widmer, "Towards explainable music emotion recognition: The route via mid-level features," in *Proceedings of the 20th ISMIR Conference, Delft, The Netherlands*, 2019.
- [37] A. Aljanaki and M. Soleymani, "A data-driven approach to mid-level perceptual musical feature modeling," in *Proceedings of the 19th ISMIR Conference, Paris, France*, 2018.
- [38] P. G. Hunter, E. G. Schellenberg, and S. M. Stalinski, "Liking and identifying emotionally expressive music: Age and gender differences," *Journal of Experimental Child Psychology*, 2011.
- [39] M. S. Larson, "Interactions, activities and gender in children's television commercials: A content analysis," *Journal of Broadcasting & Electronic Media*, 2001.
- [40] S. G. Kahlenberg and M. M. Hein, "Progression on nickelodeon? gender-role stereotypes in toy commercials," *Sex roles*, 2010.
- [41] E. Goffman, *Gender advertisements*. New York: Harper Colophon Books, 1976.
- [42] M. E. Verna, "The female image in children's tv commercials," *Journal of Broadcasting & Electronic Media*, vol. 19, no. 3, 1975.
- [43] K. A. Neuendorf, "Content analysis—a methodological primer for gender research," *Sex roles*, 2011.
- [44] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Perception*, 2010.
- [45] K. L. Whiteford, K. B. Schloss, N. E. Helwig, and S. E. Palmer, "Color, music, and emotion: Bach to the blues," *i-Perception*, 2018.
- [46] I. Schindler, G. Hosoya, W. Menninghaus, U. Beer-mann, V. Wagner, M. Eid, and K. R. Scherer, "Measuring aesthetic emotions: A review of the literature and a new assessment tool," *PLoS one*, 2017.
- [47] A. S. Cowen, X. Fang, D. Sauter, and D. Keltner, "What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures," *Proceedings of the National Academy of Sciences*, 2020.
- [48] R. Hennequin, A. Khelif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, 2020.
- [49] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015.
- [50] M. Won, S. Chun, and X. Serra, "Toward interpretable music tagging with self-attention," *arXiv preprint arXiv:1906.04972*, 2019.
- [51] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019*. IEEE, 2019.
- [52] S. Grollmisch, E. Cano, C. Kehling, and M. Taenzer, "Analyzing the potential of pre-trained embeddings for audio classification tasks," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.
- [53] C. Fine and E. Rush, "“Why does all the girls have to buy pink stuff?” The ethics and science of the gendered toy marketing debate," *Journal of Business Ethics*, 2018.