EVALUATING FOUNDATION MODELS ON TIMBRE-RELATED COGNITIVE TASKS

Anonymous Authors
Anonymous Affiliations
anonymous@ismir.net

ABSTRACT

Foundation models are increasingly applied to MIR tasks, yet their performance on music cognition problems remains underexplored. In this work, we investigate how state-of-the-art audio-language models and large language models (LLMs) perform on timbre-related cognitive tasks. We focus on music emotion recognition which captures listeners' perceived and induced emotions in response to instrument tones, and run additional tests on instrument recognition. We evaluate contrastive audio-language models (CLAP variants and MuQ-MuLan) in both zero-shot and probe-based settings, and compare their performance with Centaur, a recent LLM fine-tuned on human decision patterns. We further propose a novel inference pipeline that integrates CLAP descriptors as intermediate textual prompts for LLMs. Results show that LLMs, especially Centaur, outperform both zero-shot and probe-trained contrastive models, while the hybrid pipeline yields the best performance overall. Our findings suggest that combining audio-language and language-only models provides a promising direction for modelling music-related cognition, with implications for applications such as music recommendation, generation, and adaptive audio interfaces.

1. INTRODUCTION

Foundation models have been applied to music in a variety of domains, including representation learning, multimodal integration, and music generation in both symbolic and audio formats [1]. Though these models have been evaluated on various general music information retrieval tasks, they haven't yet been extensively tested on music cognition tasks. Music is a fundamentally cognitive activity [2] therefore computational models of music should be able to capture the perceptual and experiential aspects that music psychology investigates.

In this paper, we evaluate foundation models of music on timbre-related tasks, focusing on music emotion recognition as a case study. Timbre is a musical features that can be described as not pitch, rhythm, or loudness and is often described in natural language by both researchers and practitioners (e.g., audio engineers and producers). Many foundation models of music rely on audio representations such as mel spectrograms, which retain timbral information [1], making timbre a compelling target for evaluation. From a music cognition perspective, our goal is to test how well these models reproduce listeners' perceptual and affective responses, thereby moving beyond purely formal or structural accounts of music understanding.
This study addresses two research questions:

1. How well do foundation models capture human responses to timbre in emotion recognition and instrument identification tasks?

2. Does combining audio-language models with LLMs improve performance on music emotion recognition?

Our contributions are threefold:

- We conduct a systematic evaluation of language-audio contrastive-learning models on timbre-related cognitive tasks, comparing zero-shot inference with probe-based regression approaches.

- We assess the performance of LLMs on an timbre-emotion association task, and introduce a novel hybrid pipeline in which semantic descriptors are injected into LLM prompts.

- We provide empirical evidence that this hybrid approach yields the closest alignment with human responses, outperforming both traditional baselines and audio foundation models.

2. BACKGROUND

The few works that look into evaluation of foundation models on music cognition tasks have provided promising results [3, 4]. This is the line of inquiry that this work aims to extend.

In addition to the other music-related tasks for which foundation models have been employed, music question answering is a relatively new task that leverages their general music understanding. Despite the impressive performance of LLMs and other foundation models in question answering, significant drawbacks have been identified, most importantly the limited influence of the audio encoder on the answers compared to the text encoder [5]. To counteract this, we propose an alternative pipeline (see figure

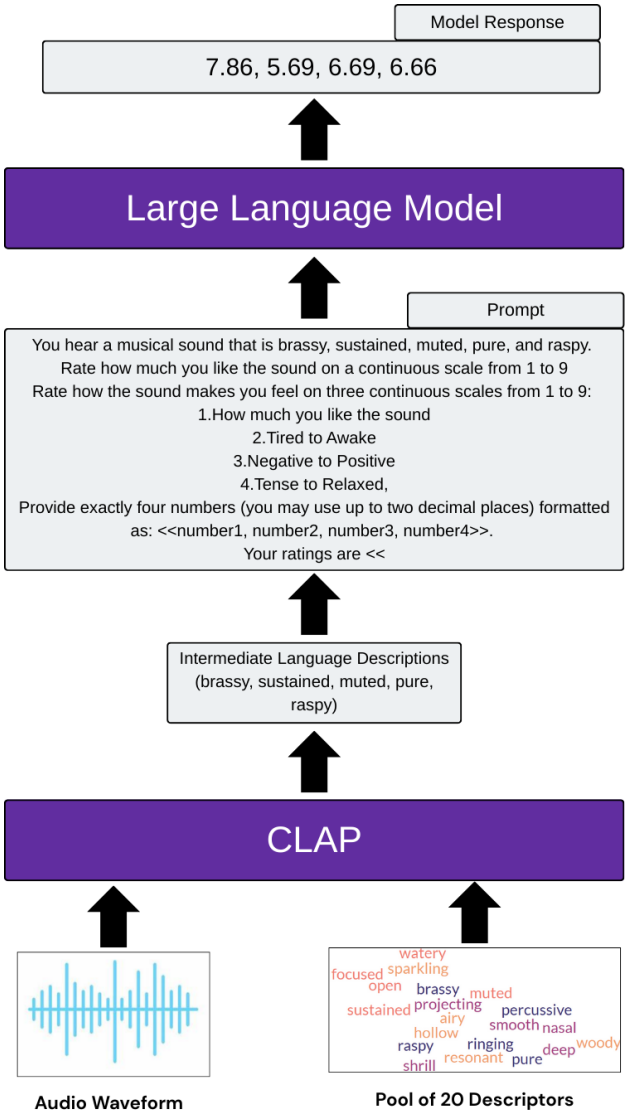

Figure 1. Proposed pipeline: CLAP extracts descriptors from a pool of 20 used in Reymore et al. [6]. These are then included in the prompt to Centaur.

8) where an audio-language model processes audio and extracts an intermediate language description, that is then fed into an LLM as part of the prompt.

3. METHOD

3.1 Data

We first focused on a dataset of human timbre-emotion associations (Experiment 1 by Korsmit et al. [7]). This consists of 59 music tones (each lasting for 3 seconds) played by 26 instruments across different types/families. Listeners (total $N = 263$) were split into four groups, each asked to report *induced* (felt) vs *perceived* (expressed in music) emotion and whether they had to choose between a *dimensional* emotion representation (consisting of valence, tension and energy plus their liking of each sample) versus a *discrete* emotion representation (consisting of five affective descriptors: happiness, sadness, anger, tenderness and fear). Therefore this dataset comprises four sub-experiments: Induced Dimensional (IDim), Perceived Dimensional (PDim), Induced Discrete (IDisc) and Perceived Discrete (PDisc).

Additionally, we performed initial tests on a recent dataset of human timbre recognition responses (Experiment 1 by McAdams et al. [8]). This includes 151 music tones, played by 11 instruments of different families, at different pitch ranges (each lasting less than a second). Listeners ($N = 25$) had to select which of the 11 instruments produced each of the 151 audio files.

3.2 Models

Two types of foundation models were used in this work: audio-language contrastive learning based models and LLMs. More specifically, CLAP [9] variants with differing audio encoder sizes and training datasets, and the recent MuQ-MuLan model [10] were used (see table 3.2 for details). In addition, Centaur (an instance of Llama 3.1 70B fine-tuned on human decisions) [11] was also evaluated, and compared to Llama 3.1 70B instruct [1], which we found to be performing much better and closer to Centaur compared to the non-instruct version.

3.3 Experiments

For the timbre-emotion association data [7], we tested the contrastive learning models using two different methods:

(1) Zero-shot inference: The audio stimuli were inputted to the audio encoder and text prompts were created and inputted to the text encoder (see section 3.4). Similarity scores were computed for each audio-text pair by taking the dot product between the audio and text embeddings. Then, the cosine similarity scores were normalised to the continuous range of 1-9, corresponding to the responses listeners gave during the experiment.

(2) MLP probe training: We also split the human responses into train and test subsets and trained MLPs to perform regression over the output values, using the audio embeddings of each of the models as input. We trained both univariate regressor MLPs (which output one emotion value at a time) as well as multivariate regressor MLPs (which output all the emotion values of either the induced or perceived category simultaneously).

For each of the above methods, we evaluated their outputs against averaged human responses using multiple regression metrics. Additionally, we converted the ratings of the models and the averaged human responses to rankings and evaluated them using multiple ranking metrics.

We used two baselines: a random one and an MLP trained on the human responses of each sub-experiment using standard timbral features used in the literature and taken from the repository provided by Korsmit et al. [7].

For the timbre recognition data [8], we run zero-shot inference using CLAP variants to test whether it could

[1] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
[2] https://github.com/LAION-AI/CLAP
[3] https://github.com/tencent-ailab/MuQ

Model Name	Audio Encoder (type/size)	Embedding size	Training data
Unfused	HTSAT	768	LAION-Audio-630K, AudioSet
Fused	HTSAT	768	LAION-Audio-630K, AudioSet
General	HTSAT	1024	Music, Speech, LAION-Audio-630K, AudioSet
Music and Speech	HTSAT	1024	Music, Speech, LAION-Audio-630K
Music	HTSAT	1024	Music, AudioSet, LAION-Audio-630K
MuQ-MuLan	MuQ	1024	Music4all [12]

Table 1. The audio-language models under evaluation. Information gathered from the github pages of CLAP [2] and MuQ-MuLan [3].

achieve the 75% pass threshold that humans had to in order to participate in the experiment.

3.4 Prompt engineering

It has been shown in previous work [13] that foundation models are very sensitive to text input. Accordingly, for the timbre-emotion association data, text inputs were either tags or captions, constructed based on the original task instructions to listeners. Specifically, tags represented the two sets of emotion axes of each sub-experiment (dimensional, discrete). Captions were constructed to further specify whether the emotion was induced/felt or perceived/expressed. Multiple variants were tested until we converged to the following captions:

- Induced: "This sound makes me feel <tag>"
- Perceived: "I perceive this sound as <tag>"

We further evaluated Centaur and Llama-3.1 Instruct using two prompt designs, both of which mirrored the instructions given to human participants, but differed in the type of information provided about each musical tone. The first prompt included only metadata, the instrument and pitch range, so the models' outputs were not influenced by audio-derived information. The second prompt incorporated five semantic descriptors, characterising each tone selected from a pool of 20 descriptors used in Reymore et al. [6]. These were chosen as the top-5 most similar descriptors according to the unfused CLAP model, which we selected for its consistently strong performance without reliance on larger audio encoders or music/speech-specific training. In this way, the second prompt conveyed timbral information indirectly, via CLAP's semantic processing.

This approach represents a novel combination of foundation models: rather than projecting audio embeddings directly into an LLM, we first extract intermediate textual descriptors with CLAP and then embed them within the LLM's prompt (Figure 8). This design allows the LLM to leverage cognitively meaningful, language-based descriptions of timbre while retaining the benefits of CLAP's audio understanding.

4. RESULTS

4.1 Audio-language models in emotion recognition

Across all sub-experiments, contrastive audio-language models outperformed the random baseline in the zero-shot setting, with the exception of CLAP Music (Figure 2). This finding is notable given the musical nature of the task: models trained on broader and more heterogeneous data (e.g., CLAP General or CLAP Music and Speech) were better aligned with listeners' judgments than those trained exclusively on music. This suggests that diversity of training data enhances generalisation, even for timbre-specific tasks.

When training MLP probes on top of audio embeddings, performance improved consistently across models. R^2 values were relatively stable (around 0.2), indicating that the audio encoders reliably captured features relevant to listeners' affective ratings. The lowest R^2 was observed for CLAP Music (0.167), again confirming the limitations of music-only training. As shown in Figure 2, probe-based results (dark blue bars) generally reduced MAE compared to zero-shot inference (light blue bars).

Interestingly, the fused CLAP model underperformed relative to unfused variants. Since all stimuli were shorter than 10 seconds, fusion of multiple audio segments provided little benefit, and may have introduced unnecessary complexity. This points to an important caveat: fusion is not advantageous for short timbral stimuli.

4.2 LLMs and hybrid pipeline in emotion recognition

Turning to the LLMs, both Llama-3.1 Instruct and Centaur, substantially outperformed all CLAP variants and MLP probes in terms of MAE (orange and green bars in Figure 2). Moreover, the proposed hybrid pipeline, where CLAP descriptors were inserted into the LLM prompts, consistently outperformed the version using only instrument and pitch range information. This demonstrates that intermediate semantic descriptions derived from audio-language models can serve as more effective cues for LLMs than symbolic metadata.

Between the two LLMs, Centaur achieved the best performance overall, outperforming Llama-3.1 Instruct in both prompt configurations. This supports the hypothesis that Centaur's fine-tuning on human decision patterns makes it particularly well-suited for music cognition tasks.

4.3 Instrument recognition

In modeling this experiment, CLAP general was used, as it was one of the best performing variants, and also showed promising results: on a subset of C4 tones across instruments, it achieved 82% accuracy in instrument recogni-

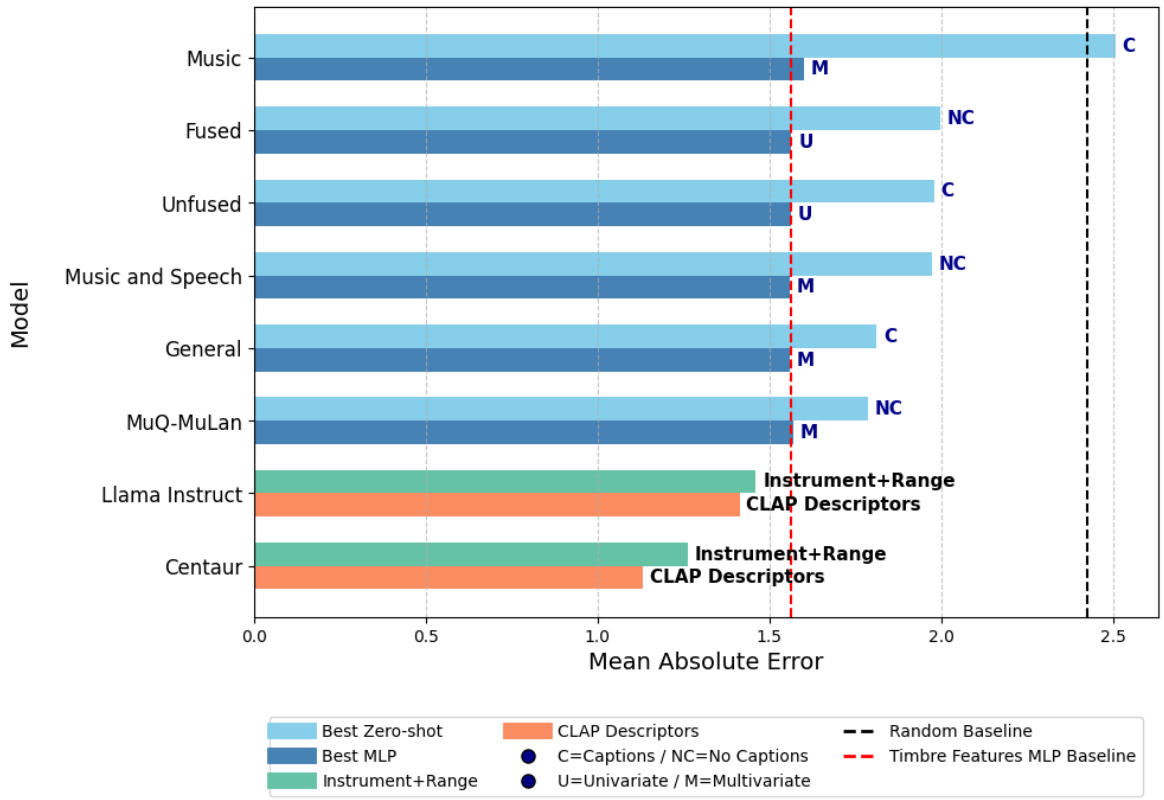

Figure 2. Overall comparison of the best variants per method across models in terms of MAE.

tion without any task-specific training, exceeding the 75% minimum accuracy that was required of human participants. This highlights the potential of foundation models to generalise beyond emotion recognition into related timbre-based cognitive tasks.

5. CONCLUSION

We evaluated foundation models on timbre-related cognitive tasks, focusing on emotion recognition and instrument identification. Contrastive audio-language models captured some aspects of human responses, but LLMs, particularly Centaur, achieved superior accuracy. Our proposed hybrid pipeline that integrates CLAP-generated descriptors into LLM prompts further improved performance, suggesting a promising strategy for bridging perceptual and linguistic aspects of music. Timbre is among the most challenging musical features to model—arguably more so than pitch, rhythm, or loudness—because it is more difficult to represent computationally, yet the tested models handled the tasks impressively well.

Future work will extend these experiments to other cognition-oriented tasks, evaluate additional music-specific LLMs (e.g. MusiLingo [14]), and explore personalisation through individual listener data. These directions could help utilise foundation models in a way that more closely reflects the cognitive and experiential dimensions of music listening.

6. REFERENCES

[1] Y. Ma, A. Øland, A. Ragni, B. M. Del Sette, C. Saitis, C. Donahue, C. Lin, C. Plachouras, E. Benetos, E. Shatri *et al.*, "Foundation models for music: A survey," *arXiv preprint arXiv:2408.14340*, 2024.

[2] G. A. Wiggins, "Semantic gap?? schemantic schmap!! methodological considerations in the scientific study of music," in *2009 11th IEEE International Symposium on Multimedia*. IEEE, 2009, pp. 477–482.

[3] R. Marjieh, I. Sucholutsky, P. van Rijn, N. Jacoby, and T. L. Griffiths, "Large language models predict human sensory judgments across six modalities," *Scientific Reports*, vol. 14, no. 1, p. 21445, 2024.

[4] K. Siedenburg and C. Saitis, "The language of sounds unheard: Exploring musical timbre semantics of large language models," *arXiv preprint arXiv:2304.07830*, 2023.

[5] B. Weck, I. Manco, E. Benetos, E. Quinton, G. Fazekas, and D. Bogdanov, "Muchomusic: Evaluating music understanding in multimodal audio-language models," *arXiv preprint arXiv:2408.01337*, 2024.

[6] L. Reymore, J. Noble, C. Saitis, C. Traube, and Z. Wallmark, "Timbre semantic associations vary both between and within instruments: An empirical study

incorporating register and pitch height," *Music Perception: An Interdisciplinary Journal*, vol. 40, no. 3, pp. 253–274, 2023.

[7] I. R. Korsmit, M. Montrey, A. Y. T. Wong-Min, and S. McAdams, "A comparison of dimensional and discrete models for the representation of perceived and induced affect in response to short musical sounds," *Frontiers in Psychology*, vol. 14, p. 1287334, 2023.

[8] S. McAdams, E. Thoret, G. Wang, and M. Montrey, "Timbral cues for learning to generalize musical instrument identity across pitch register," *The Journal of the Acoustical Society of America*, vol. 153, no. 2, pp. 797–811, 2023.

[9] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, "Muq: Self-supervised music representation learning with mel residual vector quantization," *arXiv preprint arXiv:2501.01108*, 2025.

[11] M. Binz, E. Akata, M. Bethge, F. Brändle, F. Callaway, J. Coda-Forno, P. Dayan, C. Demircan, M. K. Eckstein, N. Éltető *et al.*, "A foundation model to predict and capture human cognition," *Nature*, pp. 1–8, 2025.

[12] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues *et al.*, "Music4all: A new music database and its applications," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2020, pp. 399–404.

[13] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," *arXiv preprint arXiv:2208.12208*, 2022.

[14] Z. Deng, Y. Ma, Y. Liu, R. Guo, G. Zhang, W. Chen, W. Huang, and E. Benetos, "Musilingo: Bridging music and text with pre-trained language models for music captioning and query response," *arXiv preprint arXiv:2309.08730*, 2023.

7. APPENDIX

7.1 Additional results for Korsmit et al.

The provided figures (3-6) offer a detailed view into the performance of different foundation models on the timbre-related emotion recognition tasks. A thorough analysis of the results reveals several key findings.

The graphs illustrate that while a zero-shot approach with foundation models is a good starting point, fine-tuning with a probe MLP significantly boosts performance in both rating and ranking tasks. This is likely because the MLP learns to better map the general-purpose audio embeddings to the specific, nuanced human emotional responses in the dataset. The use of captions also provides a small but consistent improvement over simple tags in the zero-shot setting. Finally, the graphs highlight that the foundation models are capable of learning representations that are as effective as, or in some cases even more effective than those derived from traditional handcrafted timbral features, validating their use for these cognitive tasks.

- **Tags vs. Captions**: The performance difference between models using simple tags versus full captions is not uniform. While captions generally provide a slight advantage by adding context, the effect varies significantly between the two displayed models (as well as the other CLAP variants), suggesting some are more sensitive to prompt nuance than others.
- **Rankings and MLP Performance**: MLP probes, trained for regression, did not consistently outperform zero-shot models on the ranking task. This suggests that the zero-shot approach, leveraging the models' pre-trained semantic understanding, is sometimes better at capturing the relative order of human preferences than a fine-tuned regression model.
- **Dimensional vs. Discrete Emotions**: The models generally found the dimensional emotion sub-experiments (IDim, PDim) easier to predict than the discrete emotion sub-experiments (IDisc, PDisc). The continuous nature of the dimensional axes may be better aligned with the latent space of the foundation models, leading to more accurate predictions.
- **Perceived vs. Induced Emotions**: Performance was consistently higher for perceived emotion tasks compared to induced emotion tasks. This could be a direct consequence of the models' training data (e.g., AudioSet, LAION-Audio), which often contains objective descriptions of sounds rather than subjective, induced feelings.

7.2 Additional results for McAdams et al.

We further evaluated the zero-shot performance of CLAP on an instrument recognition task from McAdams et al. [8], using confusion matrices to compare its predictions against both the objective ground truth and human consensus data. The total accuracy was 63.58% when measured against the objective labels, but dropped to 56.95% when measured against the human-aligned labels. This divergence highlights a crucial finding: the model's internal representations are better aligned with the objective source of the sound than with the patterns of human perception and confusion.

The matrices revealed both strengths and weaknesses of the model. On one hand, CLAP demonstrated high-confidence recognition for instruments with distinct timbres, such as the Guitar, Tuba, and Marimba. On the other hand, it exhibited common confusions that are known to

also challenge human listeners, such as distinguishing between the Clarinet and Saxophone, or the Vibraphone and Tubular Bells. The model's difficulty with instruments like the Tubular Bells, which are likely underrepresented in its training data, further suggests that data distribution plays a significant role in performance.

To explore this discrepancy, we trained an MLP on human-aligned data from the C4 subset of tones. The MLP achieved an impressive accuracy of 90.91% on this specific subset. However, its performance on the rest of the dataset dropped dramatically to just 6.62%. This result suggests that the MLP overfitted to the unique characteristics of the C4 tones and failed to generalise effectively, unlike human participants. This massive drop in performance also points to the potential importance of the text encoder, which was not used in this probe-based approach.

This finding underscores the potential of foundation models for modeling objective acoustic properties. However, their ability to model human perception, which is a distinct and potentially more useful ability for downstream tasks such as music recommendation, requires either fine-tuning on human-centric data or a more sophisticated approach that accounts for the subjective nature of auditory cognition.

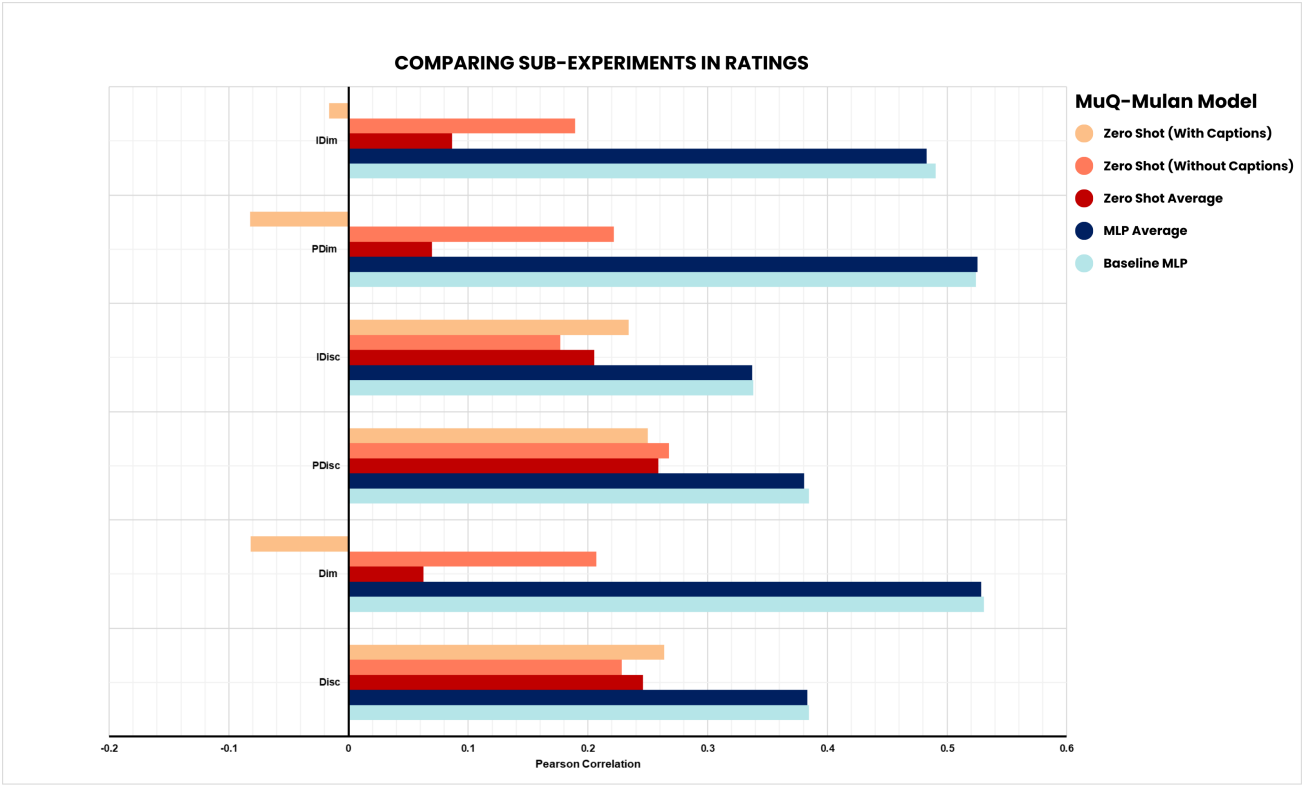

Figure 3.

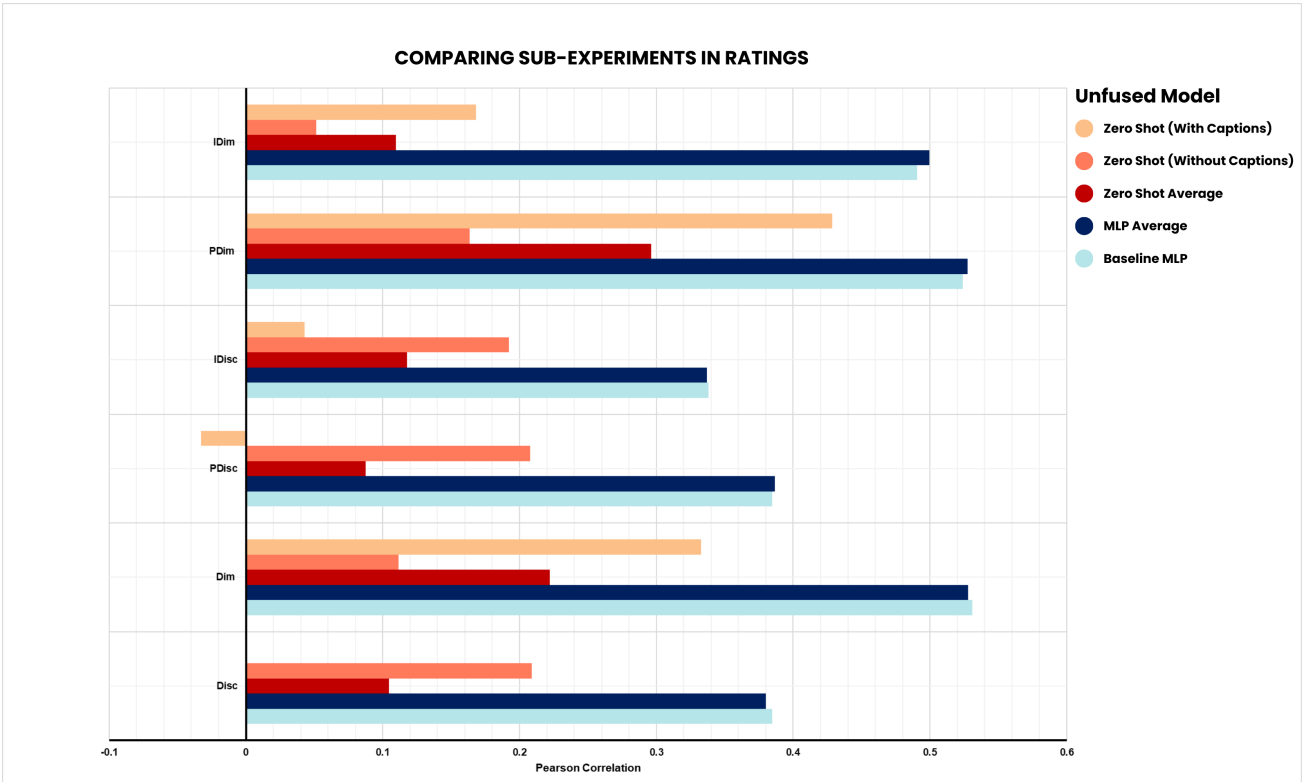

Figure 4.

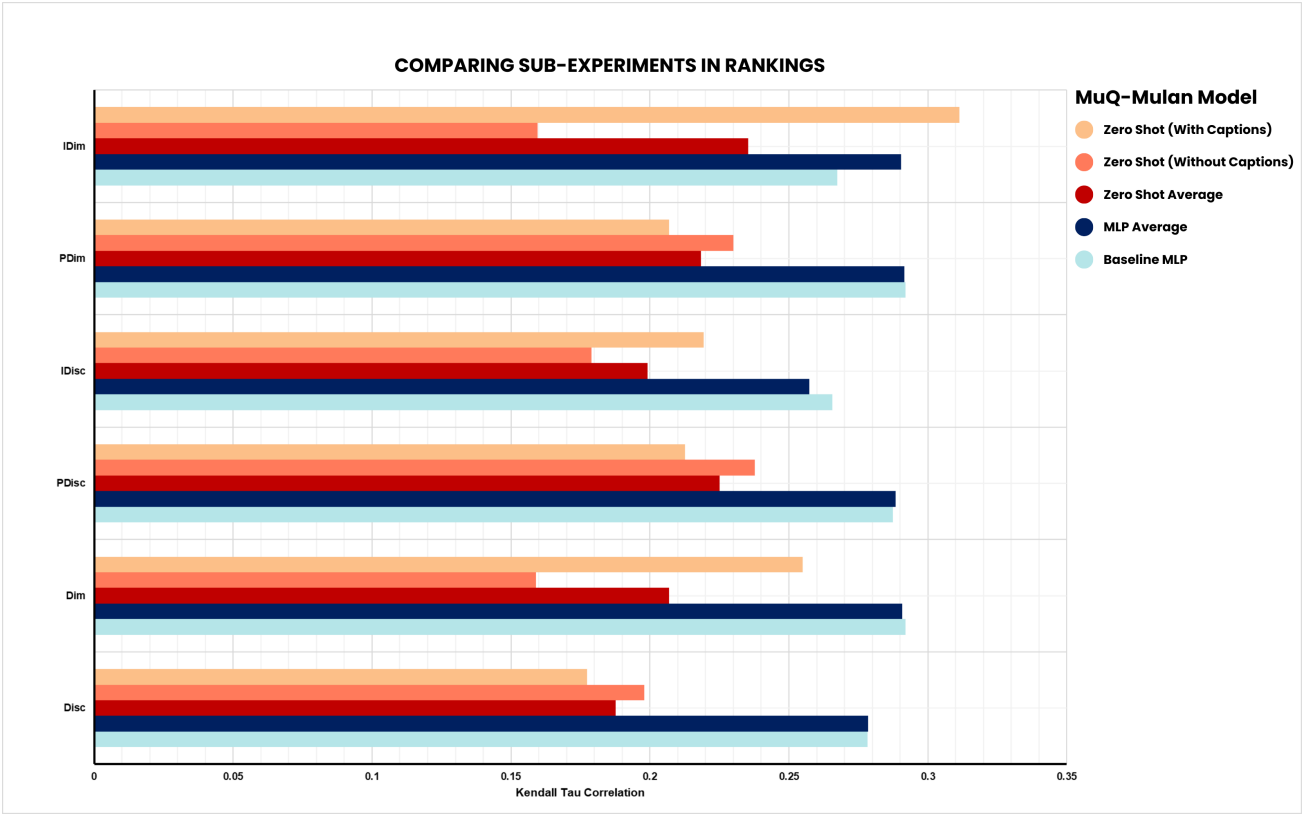

Figure 5.

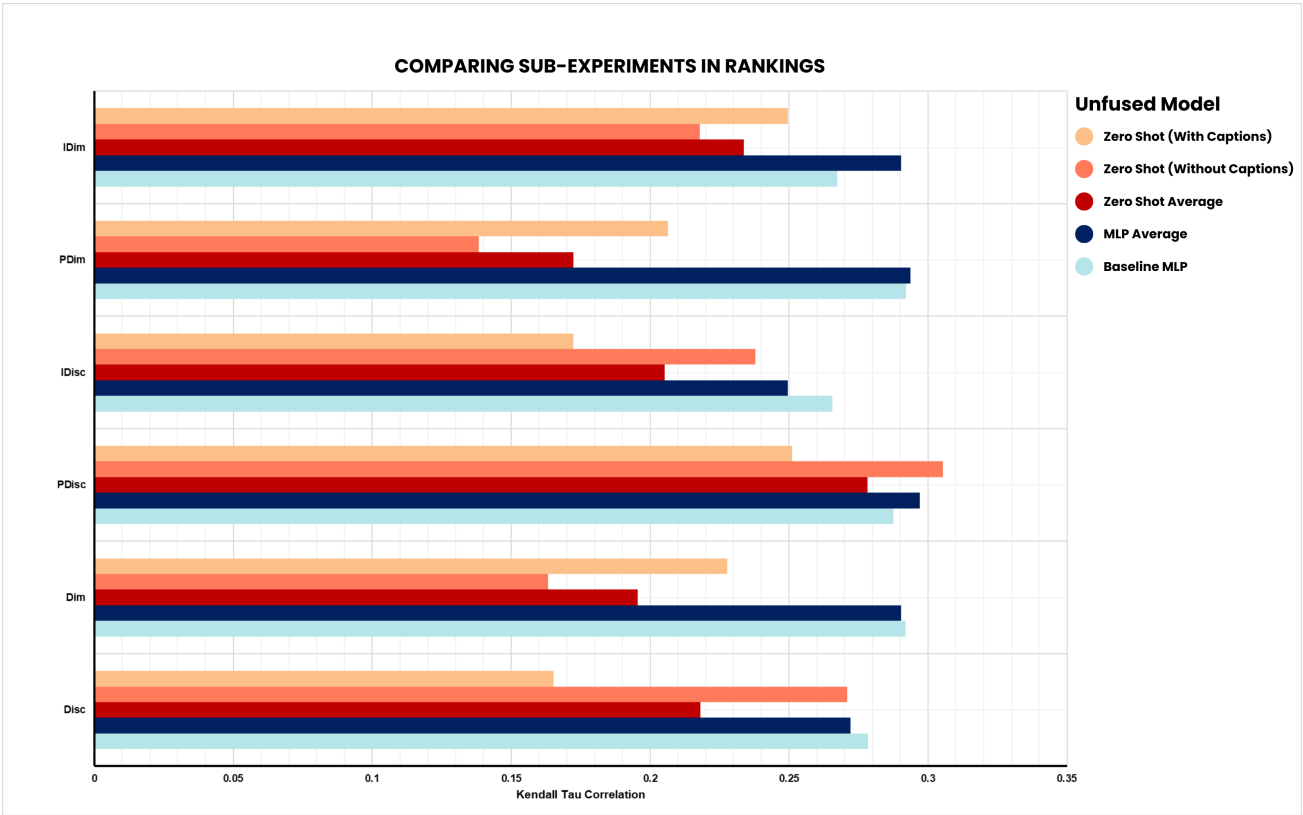

Figure 6.

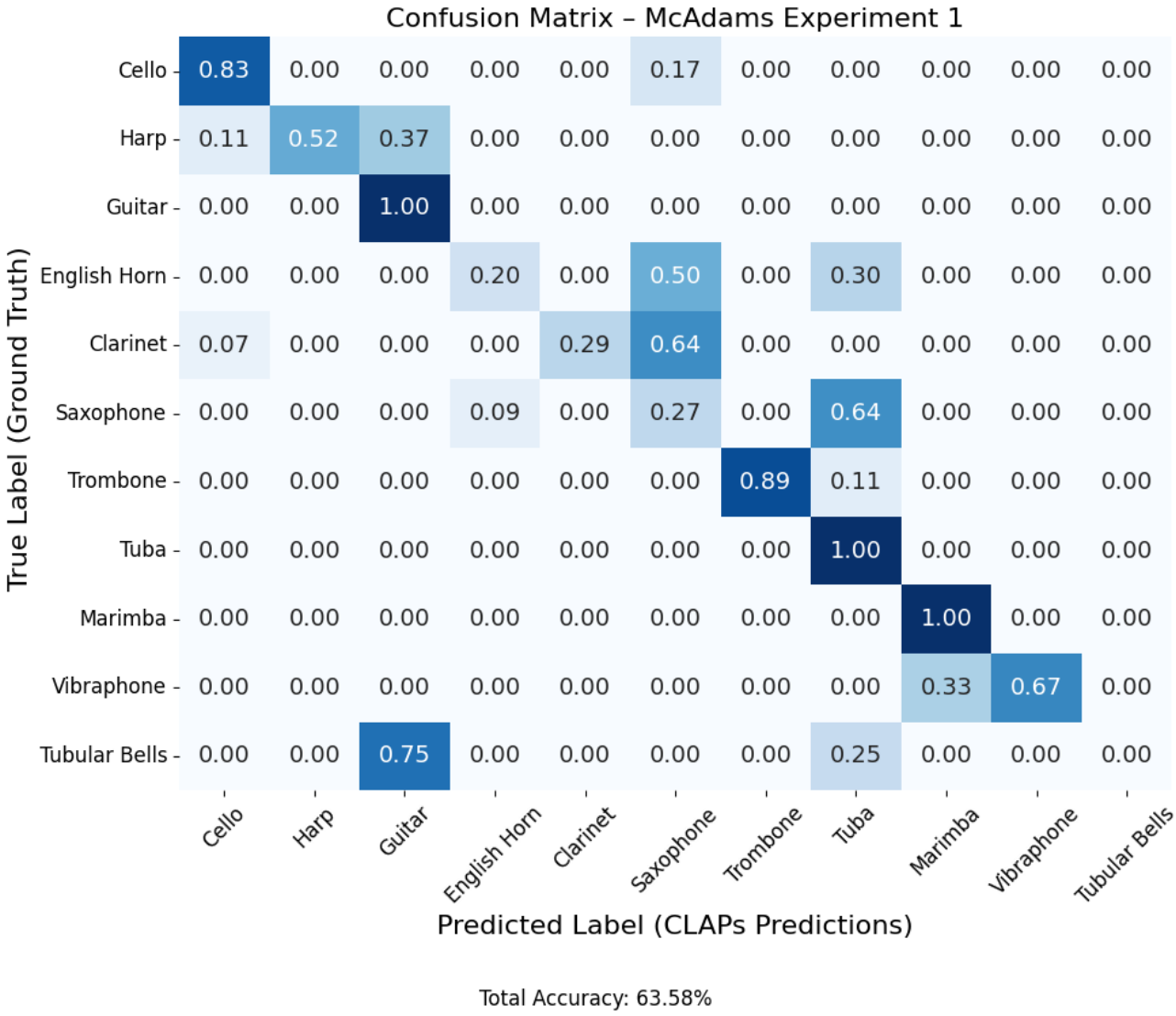
Confusion Matrix – McAdams Experiment 1
True Label (Ground Truth)
Cello 0.83 0.00 0.00 0.00 0.00 0.17 0.00 0.00 0.00 0.00 0.00
Harp 0.11 0.52 0.37 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
Guitar 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
English Horn 0.00 0.00 0.00 0.20 0.00 0.50 0.00 0.30 0.00 0.00 0.00
Clarinet 0.07 0.00 0.00 0.00 0.29 0.64 0.00 0.00 0.00 0.00 0.00
Saxophone 0.00 0.00 0.00 0.09 0.00 0.27 0.00 0.64 0.00 0.00 0.00
Trombone 0.00 0.00 0.00 0.00 0.00 0.00 0.89 0.11 0.00 0.00 0.00
Tuba 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00
Marimba 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00
Vibraphone 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.33 0.67 0.00
Tubular Bells 0.00 0.00 0.75 0.00 0.00 0.00 0.00 0.25 0.00 0.00 0.00
Cello Harp Guitar English Horn Clarinet Saxophone Trombone Tuba Marimba Vibraphone Tubular Bells
Predicted Label (CLAPs Predictions)
Total Accuracy: 63.58%

Figure 7.

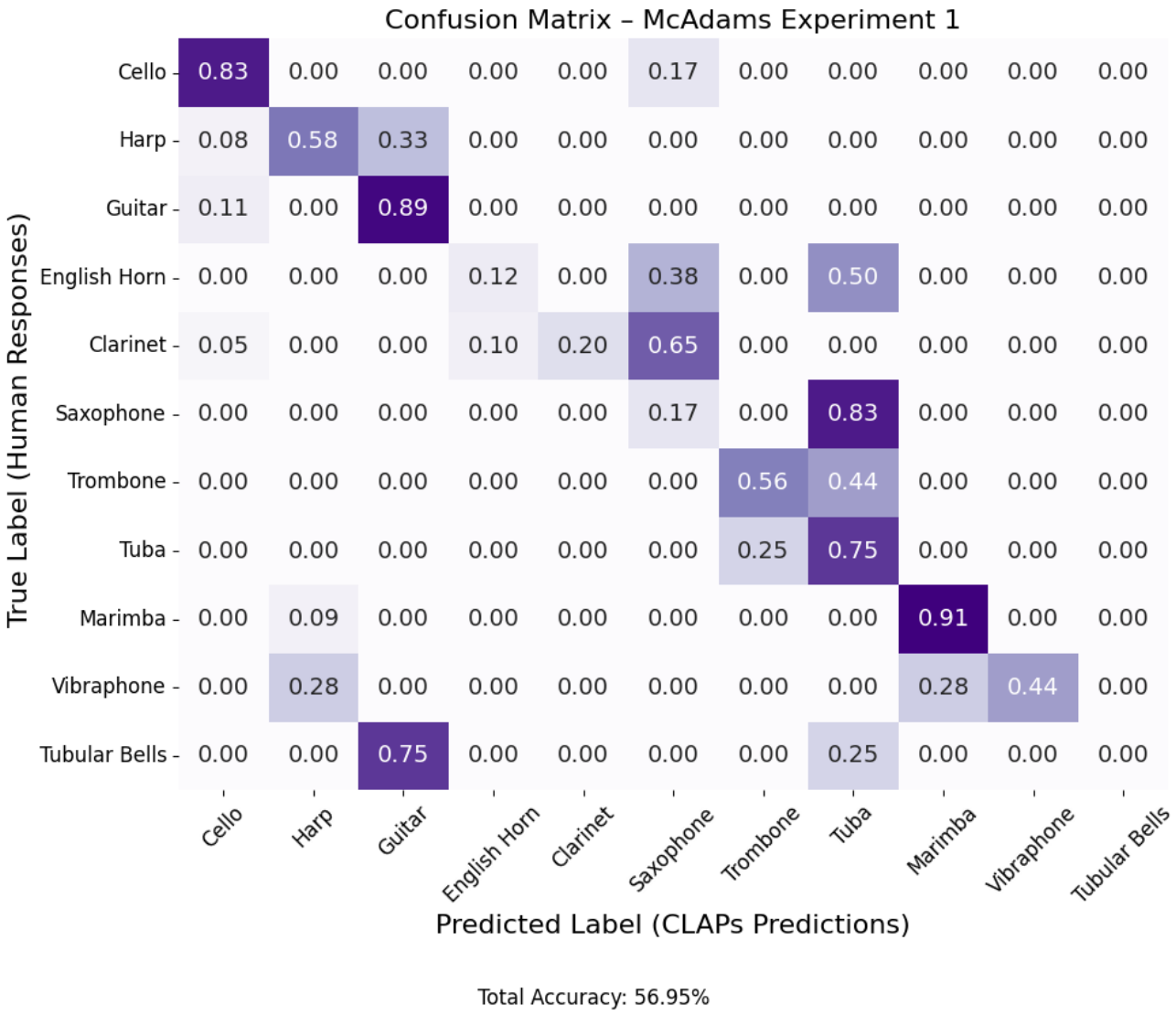
Confusion Matrix – McAdams Experiment 1
True Label (Human Responses)
Predicted Label (CLAPs Predictions)
Cello
Harp
Guitar
English Horn
Clarinet
Saxophone
Trombone
Tuba
Marimba
Vibraphone
Tubular Bells
0.83 0.00 0.00 0.00 0.00 0.17 0.00 0.00 0.00 0.00 0.00
0.08 0.58 0.33 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.11 0.00 0.89 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.12 0.00 0.38 0.00 0.50 0.00 0.00 0.00
0.05 0.00 0.00 0.10 0.20 0.65 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.17 0.00 0.83 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.56 0.44 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.25 0.75 0.00 0.00 0.00
0.00 0.09 0.00 0.00 0.00 0.00 0.00 0.00 0.91 0.00 0.00
0.00 0.28 0.00 0.00 0.00 0.00 0.00 0.00 0.28 0.44 0.00
0.00 0.00 0.75 0.00 0.00 0.00 0.00 0.25 0.00 0.00 0.00
Total Accuracy: 56.95%

Figure 8.