Models trained on procedurally generated stimuli predict human judgments of Music Acoustic Features in real-world music

Maya B. Flannery[1,3], Lauren Fink[1,3], and Matthew H. Woolhouse[1,2,3]

[1]Department of Psychology, Neuroscience & Behaviour; [2]School of the Arts; [3]McMaster University, Canada

BACKGROUND

Problem: music is difficult to describe objectively.

- Current methods require experts' (subjective) judgements.
- Commonly used methods, like genre, are ambiguous and imprecise[1].
- In some literature, music descriptions are nonexistent.

Relevance: individuals are uniquely affected by music.

- There are a wide range of music preferences and listening behaviours.
- Preferred music is beneficial to individual well-being and in therapy.
- It is difficult to investigate benefits without reliable descriptions of music.

Approach: describe music automatically by its generative features.

- Investigate compositional and performance techniques used in music generation as *features* of music[2].
- Detect such features through algorithmic analyses of digital audio[3].

OBJECTIVES

Aim: establish *Music Acoustic Features* (MAFs) as a reliable method of music classification and description for experimental research.

- **Manipulation:** computationally produce music (audio) with varying levels of six MAFs (texture, register, timbre, dynamic, tempo, and articulation).
- **Measurement:** use Essentia library for audio signal analysis and machine learning to develop models (trained on produced stimuli and applied to real-world recorded music) that predict the level of each feature.
- **Perception:** ensure that listeners' subjective judgments correspond with intended manipulations and measurements.

CONCLUSIONS

Music acoustic features have been established as:

- Manipulable: 4800 labelled audio files were systematically produced with varying levels of each MAF.
- Measurable: models trained on Essentia's extracted features predict levels of each MAF.
- Perceivable: predicted levels of MAFs correspond with listeners' judgments of MAFs for real-world music.

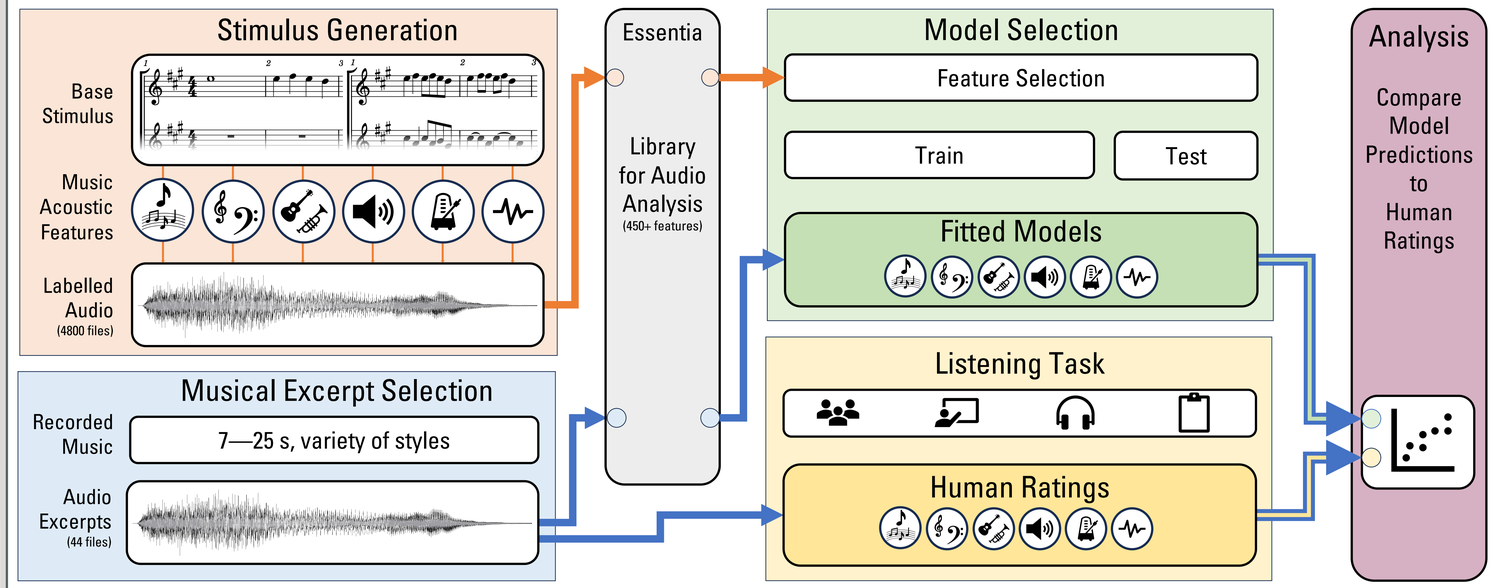
MAFs provide a consistent, objective method of music description.

- Potential to remove need for subjective judgments.
- Models can predict MAFs from audio (i.e., any existing digital audio recording).
- Next steps should expand and diversify training dataset and simplify listening task for participants.

MAFs provide a method to reliably learn how music affects individuals.

- MAFs are based on generative music features allowing for precise manipulation of stimuli in experimental studies.

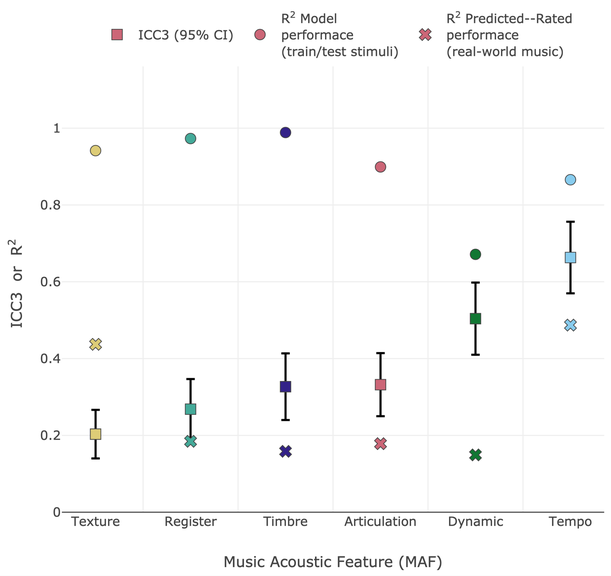
METHODS



Stimulus Generation: a seven-measure polyphonic base stimulus was written in MuseScore, then systematically manipulated by varying levels of six MAFs producing 4800 labelled audio files. **Essentia:** contains many algorithms that extract low-level, mid-level, and high-level features from digital audio. **Model selection:** potential models and hyperparameter combinations were tested to best predict MAFs from labelled audio files. **Musical excerpt selection:** 7—25 second excerpts from real-world recorded music were selected from a variety of styles. **Listening task:** listeners (N = 43) were trained to identify MAFs then listened to recorded music and asked to rate each MAF. **Analysis:** model predictions of recorded music were compared to listener ratings.

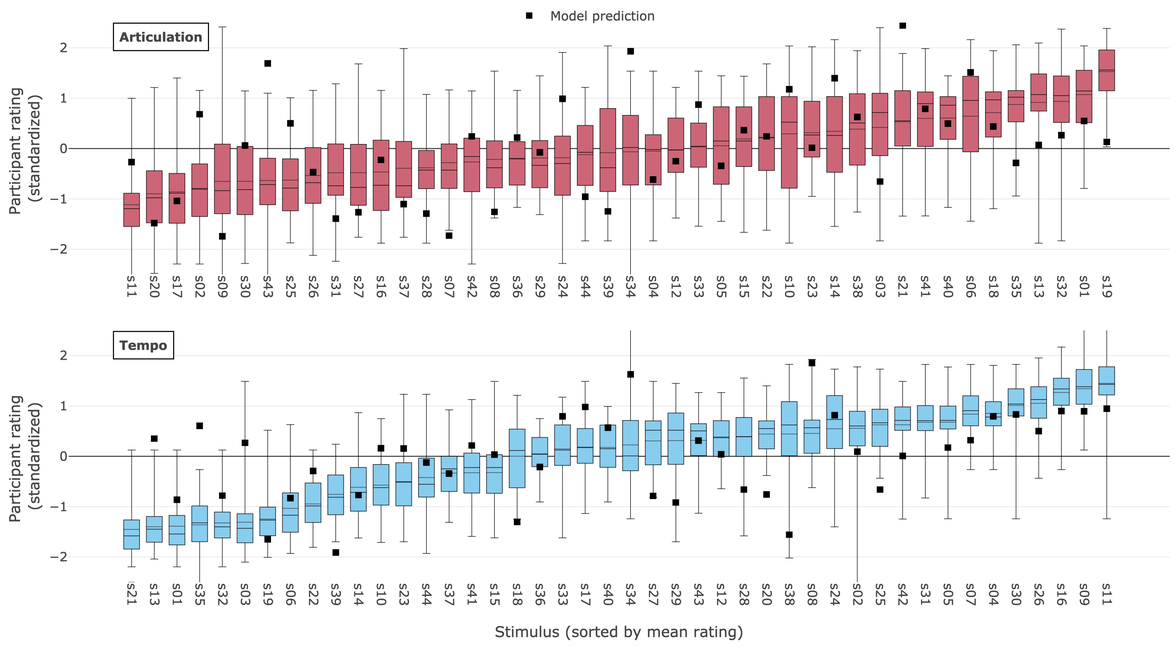
RESULTS

Participants show low to moderate MAF rating agreement across stimuli



ICC Plot: Participants listened to 44 real-world audio excerpts. The intraclass correlation coefficient (square marker) shows how well listeners agreed on each MAF level. Agreement was low to medium for most features. The R^2 for model training (circle marker) shows that models performed well on training data. However, the models performed poorly when predicting out of sample compared to listeners (cross marker).

Model predictions approximate participant responses for each stimulus within each MAF



Model—listener comparison: Two MAFs are shown above. Participant ratings are standardized and shown for each stimulus. Articulation, which had a low ICC value, contained more variable responses, and thus model predictions were less accurate (R^2 = 0.17). In contrast, Tempo had a medium ICC value and responses were more consistent for each stimulus. The Tempo model was better able to predict participant responses (R^2 = 0.48). **Limitations:** 1) Generated stimuli are not diverse enough for models to generalize well to real-world music, the training dataset should be expanded; and 2) Low agreement among listeners could be due to the listening task design (it was too difficult), the task could be simplified by reducing the number of MAF ratings per trial.

References: 1. Aucouturier, J. J., & Pachet, F. (2003). Representing Musical Genre: A State of the Art. Journal of New Music Research, 32(1), 83–93. **2.** Eerola et al. (2013). Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. Frontiers in Psychology, 4, 487. **3.** Bogdanov et al. (2013). Essentia: an open-source library for sound and music analysis. Proceedings of the 21st ACM International Conference on Multimedia, 855–858.

Visit Website

Contact: flannerm@mcmaster.ca

Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada